Title

Targeted machine learning approaches for leveraging data in resource-constrained settings

Author

Benitez, Alejandra

Publication Date

2020

Peer reviewed|Thesis/dissertation

Targeted machine learning approaches for leveraging data in resource-constrained settings

by

Alejandra Benitez

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Maya Petersen, Chair
Professor Mark van der Laan
Professor Sandra McCoy
Professor Laura Balzer

Summer 2020

Targeted machine learning approaches for leveraging data in resource-constrained settings

Abstract

Targeted machine learning approaches for leveraging data in resource-constrained settings

by

Alejandra Benitez

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Maya Petersen, Chair

Researchers in the field of public and global health continue seeking ways to reduce the disproportionate burden of disease on marginalized communities and in resource-constrained settings, for example, in low-middle income countries (LMIC). While progress towards this goal has been made by increasing the uptake of evidence-based practices (EBP) in LMIC, many barriers to sustainable implementation of EBP in LMIC remain. This dissertation is comprised of three studies which harness data-adaptive methods as a tool for supporting uptake of EBP in LMIC.

The first study sought to inform allocation of viral load tests by proposing a differentiated care approach for persons living with HIV. Our results indicate that, in comparison to current non-targeted approaches, a hypothetical machine learning approach may reduce testing frequency relative to resource-rich settings while obtaining similar sensitivity, and while maintaining delay time to viremia detection low. Relative to WHO viral load testing standards, this approach greatly improved the sensitivity and delay time to detection of viremia. The second study sought to determine whether easily attainable maternal-infant characteristics can predict risk of preterm birth, while maintaining accuracy similar to that of more expensive gestational age (GA) dating methods. Our results indicate that, among women who entered antenatal care (ANC) in the first trimester, an algorithm based on simple maternal-infant characteristics can predict GA and preterm risk within a clinically valuable margin of error. Therefore, this has the potential to inform clinical care surrounding the time of delivery and can be used for preterm birth rate reporting. The last study sought to compare methods for estimating intervention effects in small-scale cluster randomized trials (CRT), which are commonly used, particularly in resource-constrained settings. We assessed the robustness of various methods in analyzing hierarchical data, both in simulation and using a real-data example. We concluded that analyses of small-sample CRT require careful consideration surrounding weighting scheme, covariate adjustment, and target parameter specification.

To Germán and Teresa, and to our activists.

Contents

List of Figures

List of Tables

Acknowledgments

Thanks to Maya Petersen for her generous support, encouragement, and wisdom. She is a remarkable and brilliant human and mentor, and I am so lucky to have worked with her. To Laura Balzer, who has been so generous with her time and energy, thanks for getting me across the finish line. A big thanks to Mark van der Laan and Sandra McCoy, for helping their students think critically.

Thanks to Gentry, Olivia, Bill, Priscah, Kohana, Carly, Shalika, Courtney, Dana, and Jeremy, who have inspired me in more ways than they know. I am so grateful to the women of 48th St.: Eva, Beverly, and Emily, who have motivated and supported me during such formative years, and have been a constant source of thoughtful conversations and baked goods. To Lina, whose *amistad*, understanding, and great taste in music have meant the world to me. I am so proud to call her my friend.

To Mamá, Papá, Miriana, y Melissa: they have shown me what it means to be persistent, resilient, *y Mexicana*. Their sacrifice has influenced me beyond measure. To Paris and Jahel, whom I deeply admire; I am so proud of our family.

To the amazing Sharon Norris, Janene Martinez, and Sumaiya Elahi: the students and faculty in the Division of Biostatistics are lucky to have them. Thanks to the greater biostatistics and public health community; I feel fortunate to be part of the School of Public Health.

A big thank you to the PTBi-UCSF collaboration, for allowing me the opportunity to grow and learn within the PTBi East Africa team. It has been an honor to work with such an incredible group of people. Thank you to the UARTO collaboration, especially to Jessica Haberer for all her help in making our manuscript possible.

It would be difficult to overstate my gratitude to the women and families of the PTBi study, the participants of the UARTO Study, and the women of the Gasabo District in Kigali. My hope is for their voices to be heard, loud and clear, and that together, we continue working towards our goal of healthier, more equitable communities.

Lastly, the first half of 2020 has reminded us how much work we have to do in public health and as a society. To the voices that continue to speak for the oppressed, thank you.

Chapter 1

Introduction

1.1 Burden of disease in low-middle income countries

Public health efforts must continue to develop sustainable evidence-based practices (EBP) and inform policy in order to combat the disproportionate burden of disease on marginalized communities, specifically, on low-middle income countries (LMIC) [1–4]. We define EBP as medical interventions shown to have efficacy for individual-level health outcomes. However, many EBP are often not feasible to implement in LMIC, or are used sub-optimally due to several factors, such as limitations in infrastructure and data systems [5–11]. For example, in the field of HIV research, antiretroviral therapy (ART) is known to effectively treat persons living with HIV, but many barriers in LMIC limit optimal adherence to ART regimen [12]. A way forward in these settings is to identify which persons require differentiated care (treatment guidelines adapted to patient characteristics) to improve their regimen adherence [13, 14]. A second example is the burden of preterm birth in LMIC. Low-cost EBP to improve neonatal outcomes are available but are not routinely used [15, 16]. Additionally, preterm birth identification is particularly challenging in LMIC, and complicates the accurate provision of EBP to preterm infants. These challenges highlight the importance of tailoring EBP to community stakeholder needs so that they may be sustainable in LMIC.

While several pathways are used in global health research to improve EBP uptake in LMIC, in this dissertation, we consider two specific approaches. First, we demonstrate two case studies which leverage available data to inform resource allocation surrounding EBP. Secondly, we consider small-scale cluster randomized trials (CRT), which are often used to assess how individual-level EBP may scale up in the community. In both cases, recent advances in machine learning offer a means to improve both the delivery of interventions and the analysis of studies designed to evaluate effectiveness of EBP at scale. We consider these in the context of two public health crises which both disproportionately burden LMIC: the HIV epidemic, and neonatal mortality.

In this chapter, we first provide a brief overview of relevant machine learning and data-adaptive approaches with a focus on ensemble learning and loss-based estimation. Then we

discuss how these approaches can be used for data-driven resource allocation and precision improvement in small-scale CRT, which are the focus of this dissertation. Lastly, we give a brief overview of the studies and stakeholders which made this dissertation possible.

1.2 Overview of estimation using supervised learning

Substantial developments in statistical estimation methods, combined with rapidly growing data sources and computing capacity, have proliferated use of machine learning techniques across various disciplines [17, 18]. Researchers now have a wide variety of options for estimation problems. For example, to estimate the conditional risk of an event $\mathbb{E}[Y|W]$ (for binary outcome Y given covariate set W) it is unclear which method is superior. In this dissertation, we consider data-adaptive methods which automate the process of selecting the optimal algorithm with respect to a pre-specificed criteria. This general approach serves to remove user biases in the selection of the model [19].

More specifically we consider loss-based learning, in which the target estimand Ψ is explicitly defined as the minimizer of the expected loss, or risk, with respect to the underlying, unknown distribution of the data, P. This is denoted $\Psi = arg\min_{\psi} \mathbb{E}_P L(\psi, O)$. Here, O denotes the observed data, and L is the function which assigns a measure of performance to a candidate function ψ when applied to each independent unit of O [19, 20].

Further, using cross-validation, we can obtain an honest estimate of the risk for each candidate estimator $\hat{\Psi}$ based on our data. Cross-validation creates a random split of the n i.i.d. observations into m folds. The 'training' set is defined as the data consisting of all folds except for the m-th validation fold. We use $P^v_{n,m}$ and $P^t_{n,m}$ to denote the m-th empirical distribution of the validation and training sets, respectively. For each candidate estimator $\hat{\Psi}$ of our target parameter, using our loss function, we estimate the cross-validated risk and choose the candidate estimator $\hat{\Psi}$ which minimizes the cross-validated risk. $\hat{\Psi}(P^t_{n,m})$ denotes the estimator fit on the m^{th} training set, and this fit is then used to predict the outcome of interest in the validation fold. The pre-specified loss function is applied to each prediction in the validation fold to assess performance, denoted: $P^v_{n,m} L(\hat{\Psi}(P^t_{n,m}))$ (using notation $Pf \equiv \mathbb{E}_P f$ for some function of observed data f). This is done for each of the m folds, and once the risk has been evaluated in each validation set, the risk is averaged across m folds to obtain a cross-validated risk estimate [19]. The cross-validated risk for a candidate estimator $\hat{\Psi}$ is denoted

$$\frac{1}{M} \sum_{m=1}^{M} P^v_{n,m} L(\hat{\Psi}(P^t_{n,m}))$$

Throughout this dissertation, we leverage a type of loss-based learning method which incorporates cross-validated risk to build the prediction algorithm. We focus on a stacking algorithm known as super learner [21]. The general stacking approach is based on an ensemble of candidate prediction algorithms, whose combined predictions generally perform better than any single one of the algorithms considered [22]. If we choose the single algorithm in

the ensemble with the best cross-validated risk, this is shown to recover theoretically optimal properties [23, 24] and is known as "discrete super learner." Alternatively, we can combine the ensemble of predictions using a weighted average to yield the super learner's predictions, to potentially improve performance [19]. Once we have built the the super learner, the algorithm has utilized all the data. Therefore, to obtain honest performance metric estimates of the super learner algorithm itself on independent data, the recommended approach is to use the cross-validated super learner [19], which adds an external layer of cross validation.

Lastly, the super learner algorithm allows the user to flexibly choose a loss function from a large set of options. This allows the statistical estimation problem to be tailored to the research question. For example, the research question of interest may seek to optimize the area under the receiver operating characteristic curve or the precision-recall curve. Alternatively, the research question may involve a constrained loss function [25], such as the sensitivity-constrained rate of negative predictions. Throughout this dissertation, we will consider several different loss functions.

1.3 Machine learning to monitor outcomes and inform resource allocation

We now describe how machine learning methods may help support sustainable implementation of EBP in LMIC. Data-adaptive methods have the capacity to monitor individual-level outcomes, thereby informing targeted interventions and resource allocation in LMIC. In the context of effectively delivering ART to persons living with HIV, as described above, machine learning may help identify individuals facing ART regimen adherence challenges in order to (i) deliver differentiated care, and (ii) allocate viral load testing for those most at risk for viremia. Similarly, machine learning methods may help more accurately identify risk of preterm birth, which in turn can (i) guide clinical care and allocation of clinical resources, and (ii) improve reporting of preterm birth to inform policy.

Identifying higher risk individuals (in the context of a given intervention) is complex. Increased accessibility of devices which store, transfer, and analyze data has spurred rapid growth of rich information sources, which can be used to estimate risk. For example, decreasing costs of devices such as tablets and cell phones have allowed them to become more readily incorporated in research in LMIC [26, 27]. However, the resulting data are often high-dimensional, noisy, and vulnerable to sparsity and missingness [28]. Machine learning provides necessary data-adaptive qualities for analyzing these data, and helps unify various data streams to predict risk and target resources accordingly.

Several recent studies have invoked machine learning methods to estimate risk of health outcomes in LMIC [29–34], proposing the potential to inform delivery of targeted interventions and support the uptake of EBP. Despite the potential advantages surrounding risk identification using data-adaptive approaches, issues surrounding the deployment of these tools in LMIC warrant careful consideration. Implementation of these algorithms in many

regions of LMIC still presents a challenge; algorithms must be carefully adapted to each community and are usually not transferable [6]. Robust data quality and other infrastructure are required to deploy these approaches at scale, and ethical standards to protect the interests of communities in LMIC must be in place [35, 36].

In Chapters 2 and 3, we present two case studies which contribute to the literature supporting use of data-adaptive methods to aid the uptake and delivery of EBP in rural and peri-urban regions of East Africa. While these studies consider the barriers to deployment mentioned above, their main objective is to understand the utility of data-adaptive methods for tailoring targeted interventions, given the constraints in these rural and peri-urban regions.

1.4 Machine learning for precision gain in cluster randomized trials (CRT)

We now consider CRT, which are designed to measure the effect of a group-level intervention. The growing CRT literature in economics, public policy, and public health seeks to understand how interventions affect outcomes at the scale of, for example, neighborhoods, schools, or hospitals, rather than only at the individual-level [37–39]. These types of studies allow us to understand how individual-level EBP may scale up to the community-level, shedding light on the required resources and structures that need to be in place to sustain improvement of outcomes.

There exist many barriers to the implementation of CRT in LMIC. The resources required for implementation, such as infrastructure, transportation, and personal equipment [7, 10], may limit the size and scale of the CRT. Enrolling very few clusters in a CRT may be often be the most feasible approach in resource-constrained settings. This in turn limits statistical power. Therefore these constraints should be taken into account in the analysis of small CRT in particular.

It is well known that effect estimates in CRT can gain power and precision by adjusting for baseline covariates [40–42]. However, adjusted effect estimates for small-scale CRT are particularly susceptible to overfitting. In this case, loss-based learning and data-adaptive methods are useful for effect estimation, for example, through automating the selection of the adjustment set using a pre-specified criteria [42, 43]. Previous analyses have compared performance of several estimators in simulation and real data [44–46]. However, many of these methods are still not well understood especially in the context of small-scale CRT with fewer than 40 clusters. To the best of our knowledge, the study in Chapter 4 is the first CRT comparative methods analysis to include targeted maximum likelihood (TMLE) methods. Our CRT comparative methods study specifically focuses on small-scale CRT and takes into consideration limitations specific to resource-constrained settings.

1.5 Overview of studies

Several different studies which took place across rural and peri-urban East Africa are discussed in this dissertation. Below, we describe the communities involved in making this research possible, and provide an overview of the motivation for each study.

Uganda AIDS Rural Treatment Outcome (UARTO)

In Chapter 2 of this dissertation, we consider the Uganda AIDS Rural Treatment Outcome (UARTO) study (NCT01596322) which took place in Mbarara, Uganda, a city in southwest rural Uganda [47]. Participants in the study were initiating HIV ART, and received free ART through the Mbarara Regional Referral Hospital Immune Suppression Syndrome (ISS) between 2005 and 2015. ART regimen adherence data and viral load data were collected in order to understand behavioral determinants and biological consequences of incomplete adherence [47, 48]. Qualitative interviews were also conducted to understand the experience of medication adherence monitoring, in order to inform subsequent data collection [48].

Cohort study in the Gasabo district of Kigali, Rwanda

In Chapter 3, the analysis leverages data based on a longitudinal cohort study which took place in Rwanda between September and October 2017, in collaboration with the College of Health Sciences at the University of Rwanda [49]. Women seeking ANC across 10 health centers in the Gasabo district of the Kigali Province were invited to enroll in the study if they met eligibility criteria. The study was designed specifically to investigate maternal genitourinary infection and maternal micronutrient deficiency as risk factors for preterm birth. The study aimed to identify preterm birth risk factors that could be prevented or diminished through targeted interventions.

The PTBi Study

In Chapters 3 and 4 of this study, the real-data examples are based on the Preterm Birth Initiative (PTBi) cluster-randomized study. Several health facilities across Kenya, Uganda and Rwanda enrolled in the PTBi study. We will describe the two sub-studies of PTBi that took place in difference regions of East Africa.

Intervention in health facilities across Rwanda

In Chapter 3, we discuss the PTBi-Rwanda study, which was a partnership between the University of Rwanda, the Rwanda Biomedical Center, and the Ministry of Health (MOH) (NCT03154177) [50]. The primary outcome of the PTBi-Rwanda study was the mean difference in gestational age (GA) at birth between arms exposed to group versus standard antenatal care (ANC). The study was motivated by an intervention in which exposure to

group ANC was shown to improve birth outcomes in a cohort of North American women [51]. Stakeholders in Rwanda sought to implement a similar intervention in their own communities, to test whether Rwandan women receiving group ANC experienced increased GA at birth, compared to women receiving the individual model of ANC [50]. 36 health centers across Rwanda were pair-matched and, within each pair, randomized to either individual ANC (control) or group ANC (intervention). Women who initiated ANC between May 2017 and December 2018 were invited to participate if they met study inclusion criteria. Additionally, urine pregnancy test (UPT) and basic obstetric ultrasound by were offered in half of the health centers in this study, to determine whether these would increase attendance and early presentation for ANC [52].

Intervention in health facilities across Kenya and Uganda

In Chapter 4, we consider the study designed to evaluate the impact of a health facility intervention on 28-day mortality among preterm infants in Eastern Uganda and Western Kenya from October 2016 to May 2018 (NCT03112018). Twenty public sector health facilities across Western Kenya and Eastern Uganda, consisting of large hospitals and smaller health centers, were pair-matched and the randomized to either the intervention or control arm [53]. Local stakeholders together with the PTBi designed an intervention package to increase the uptake and use of EBP to improve the quality of intrapartum and newborn care [54]. Stakeholders and researchers selected the following four components, with the objective of reducing the combined rate of fresh stillbirth and neonatal mortality among preterm births [54]:

1. strengthening of routine data collection
2. a modified WHO safe childbirth checklist (tailored to better identify preterm labor at presentation to anticipate specific needs).
3. PRONTO^{TM} Simulation training [55]
4. quality improvement (QI) aimed to reinforce and optimize use of evidence-based practices

The facilities in the control arm received the first two components only, and facilities in the intervention arm received all four components. All study components consisted of known interventions aiming to holistically improve quality of care, teamwork, and data use throughout the triage, delivery, and newborn periods [53]. Women seeking care for delivery at facilities participating in the PTBi study were eligible to enroll, and therefore, these women were exposed to the intervention at the time they presented for delivery.

Chapter 2

Differentiated care approaches for viral load testing compared to non-targeted approaches

2.1 Introduction

World Health Organization (WHO) guidelines now recommend antiretroviral treatment (ART) for all persons living with HIV, the majority of whom live in resource-limited settings [56, 57]. International consensus is increasing that effectively implementing universal treatment will require a differentiated care strategy, with the intensity of clinical follow-up and monitoring varying based on individual patient need [13]. In particular, the WHO now recommends that stable patients have plasma HIV RNA levels (viral loads) monitored less frequently than the quarterly monitoring previously recommended for all patients [57]. Here, stability is defined as evidence of treatment success after receiving ART for at least 1 year [58].

While decreasing monitoring frequency for stable patients can reduce costs for treatment programs and patients, it remains unclear how to most effectively identify patients in need of more frequent monitoring. Delayed detection of adherence lapses and ongoing detectable viral replication (viremia) can harm patient health, contribute to viral resistance, and increase risk of HIV transmission [59–62]. Strategies for tailoring monitoring intensity based on evolving metrics of patient risk for viremia are needed to optimize both the impact and the cost-effectiveness of differentiated ART delivery systems [63].

Electronic adherence monitoring (EAM) systems, which record a time-date stamp whenever a medication storage device is opened as a proxy for medication ingestion, provide data to potentially inform such strategies. EAM systems could be used in combination with clinical data to identify patients at increased risk of viremia, triggering both additional viral load monitoring to detect viremia and adherence interventions to prevent it. EAM data can now be accessed in real-time through cellular networks. The costs of this technology are falling

[27, 47]; however, the extent to which EAM systems can inform differentiated care decisions remains unclear.

EAM data can be summarized with many possible adherence metrics, including the proportion of prescribed doses for which an event is recorded, timing of device openings, and duration and frequency of lapses in openings. How best to select among these metrics and combine them with clinical data (such as duration of viral suppression and pre-ART CD4+ T cell count) to assess risk of viremia is unknown. Modern machine-learning approaches address this challenge by developing flexible and complex syntheses of EAM and clinical data to predict viremia more accurately. Analysis of standard EAM data (stored on a device, but not available in real-time) from HIV patients in the United States demonstrated that machine-learning can improve prediction and classification of viremia [64]. However, this approach has yet to be evaluated using either real-time EAM data, which may differ in patient use and/or accuracy compared to standard EAM, because real-time EAM allows for real-time data quality corrections. Additionally it has yet to be evaluated in a resource-limited setting, where individual, immunological, and virologic factors may differ from resource-rich settings [65–68].

We used machine learning methods to analyze standard and real-time EAM data from an observational cohort of persons living with HIV in rural Uganda, and evaluated the added value of EAM technologies to predict viremia, beyond the information provided by standard clinical and demographic data. We further assessed the potential for real-time EAM data to effectively differentiate viral load testing frequency while minimizing delays in viremia detection.

2.2 Methods

Study Population

We analyzed data from the Uganda AIDS Rural Treatment Outcome (UARTO) study (NCT01596322), an observational cohort of adults (>18 years) living with HIV who initiated ART in Mbarara, Uganda between 2005 and 2015. Participants lived within 60km of the Mbarara Regional Referral Hospital Immune Suppression Syndrome Clinic, which provides free ART in the region. To be eligible for real-time monitoring, participants required a cellular signal at home [69].

Measures

ART adherence was monitored between 2005-2011 using standard electronic pill bottles (the Medication Event Monitoring System [MEMS], West Rock, Switzerland), from which time-date stamps recording each device opening were downloaded onto a laptop computer via a USB cable during monthly home visits. From 2011-2015, adherence was measured using a real-time electronic monitor that transmitted device opening data to a web-based server

using cellular networks (Wisepill; Wisepill Technologies, South Africa) [27]. Lapses in device openings >48 hours detected using real-time monitoring triggered a home visit to determine the cause of lapse. Participants were enrolled through 2012; thus, some participants were monitored with both device types.

Viral loads and CD4+ T cell counts were measured approximately quarterly, according to research protocol; after 2011, additional viral load measures were administered during home visits following adherence interruptions detected during real-time monitoring. Viremia was defined as a single viral load >1000 copies/ml, a threshold chosen to match WHO guidelines and minimize "blips" (temporary, low-level increases in viral load) [70].

Statistical methods

Risk Score Development

We used an ensemble machine learning method to build prediction models for viremia. Viral loads measured $\leq$ 90 days after ART initiation were included as outcomes. Because detection of viremia could affect both subsequent viral non-suppression and monitoring, viral loads occurring after first detection of viremia were censored. Risk scores were constructed separately for participants followed with standard versus real-time EAM. Several candidate predictor sets were considered (Supplement I).

1) "Clinical" predictors included age, biological sex, CD4+ T cell counts before and after ART initiation, and ART regimen (drugs, regimen changes, prescribed dosing interval, and time since initiation).

2) "EAM + Clinical" predictors augmented the clinical predictors with additional EAM data. Candidate EAM features were evaluated over a range of periods (Supplement I) preceding each viral load measurement (from 7 to 365 days). For each of these periods, we calculated daily adherence (number of EAM events/total number of prescribed doses), variance of daily adherence, minimum adherence, number and duration of interruptions in events, and variability in timing between recorded events. To evaluate the extent to which ongoing CD4+ monitoring improved prediction in the context of EAM, we also considered a predictor set excluding post-ART initiation CD4+ counts.

3) "Full EAM" predictors augmented clinical and EAM predictors with viral load data, including either a) viral load at ART initiation only or b) all time-varying viral load data.

Predictor variables missing a measurement were imputed using last measured value, with time since last measurement included as a predictor. Tests with missing predictor values (after imputation) or occurring after a >400-day lapse in testing (threshold chosen based on the distribution of lapses in monitoring) were excluded.

Super learner, an ensemble method which combines several "candidate" machine learning algorithms using internal cross-validation, was used to construct a prediction model for viremia for each predictor set [21]. Leaving aside each fold in turn as validation data, candidate prediction algorithms were fit on the remaining 9/10ths of the data. Validation data were then used to select the convex combination of algorithms that maximized the

rate of negative prediction under a constraint to maintain sensitivity above 93% (constraint raised to 95% in sensitivity analyses), together with a corresponding cutoff for positive classification [25]. This threshold was chosen to improve rate of negative prediction while maintaining a clinically acceptable sensitivity. The following algorithms were included as candidates: gradient boosting machine [71], random forests [72], Bayes generalized linear models [73], and elastic net [74], each with and without a dimension reduction based on marginal correlation with the outcome.

Performance

An additional layer of cross-validation was used to evaluate the performance of the prediction models by calculating performance metrics in each independent validation set and averaging. Individual participants were stratified based on viral suppression status before sample-splitting to ensure that each fold had a similar class balance; all sample splitting respected the individual as the unit of independence. While the machine-learning algorithm aimed to optimize differentiated testing rather than to accurately predict the full range of risk, as global measures of performance we plotted cross-validated receiver operating characteristic (ROC) curves and calculated area under the ROC curve (cvAUC). Differences in cvAUC for the subsets of variables were tested using the influence function of the cvAUC to derive a z-test [75]. We also calculated net reclassification improvement, and, for our primary EAM predictor set, plotted calibration and conducted the Hosmer-Lemeshow test (Supplement 2).

We then evaluated the potential of each of the candidate predictor sets to reduce viral load testing frequency and increase yield when combined with a selected cutoff chosen in the corresponding training set (to accurately assess performance of the learned testing rule versus the risk predictor alone on independent data). Specifically, we calculated the cross-validated rate of negative prediction (cvRNP) (proportion of viral load tests that would have been avoided because predicted risk of viremia was below cutoff), "number needed to screen" (cvNNS; number of viral load tests with predicted risk above cutoff / the number of viremia cases with predicted risk above cutoff), empirical sensitivity (cvSens; proportion of viremia cases with predicted risk above cutoff), false positive rate (cvFPR; proportion of non-viremia cases with predicted risk above cutoff), and the precision (cvPPV; number of viremia cases with risk score above cutoff / number of risk scores above cutoff).

We further evaluated cross-validated performance of three hypothetical strategies for viral load monitoring: 1) A"3-month" schedule, in which viral load was measured every three months (a reference for comparison that, while not realistic in many settings, was available for the study population); 2) A "WHO" schedule: in which viral load was measured at 6 and 12 months after ART initiation and annually thereafter, as recommended for stable patients [[58] (a schedule now routine care in Uganda); and, 3) An "EAM"-based differentiated monitoring strategy, in which the WHO schedule was augmented with additional viral load tests on dates that the predicted risk of viremia exceeded cutoff (using all predictors except viral loads and restricting EAM-triggered tests to dates without missing predictors). For

each strategy, we calculated the monitoring rate (tests ordered per person-year), sensitivity, NNS, FPR, and delay to viremia detection relative to observed date of first detection under research protocol (corresponding to maximal but unrealistic testing frequency). To do so, we assumed that omitting an observed test would not change future adherence or viremia and that if an observed viremic test were omitted, viremia would still be present and would not be detected until the next test.

Analyses were performed using R version 3.5.0 [76], and packages SuperLearner [77], xgboost [78], bartMachine [79], glmnet [80], arm [81], ROCR [82], and predictABEL [83].

Ethics

The Mbarara University of Science and Technology, the Uganda National Council for Science and Technology, Partners Healthcare, and the University of San Francisco, California ethical review boards approved this study. All participants provided written informed consent.

2.3 Results

Sample Characteristics

A total of 443 participants were monitored with standard EAM for a median of 4.6 years (IQR: 2.5-5.6) and contributed 5,922 viral load results as outcomes in the standard EAM analysis dataset; 485 participants were monitored with real-time EAM for a median of 2.2 years (IQR: 1.4-2.5) and contributed 2,834 viral load results as outcomes in the real-time EAM analysis dataset (Table 2.1). Of real-time EAM users, 243 had been monitored with standard EAM before initiating real-time monitoring. Between 2005-2011, 86 of 443 participants (19%) monitored with standard EAM experienced viremia (86/5923 tests, 1.5%). Between 2011-2015, 45 of 485 participants (9.3%) monitored in real-time experienced viremia (45/2834 tests, 1.6%).

Consecutive viral load tests were a median of 105 days (IQR: 97-114) and 115 days (IQR: 97-178) apart for standard and real-time EAM users, respectively; 4% of tests under standard EAM monitoring and 25% of tests under real-time EAM monitoring were administered $<$60 days since prior test; under real-time monitoring 36% of these were preceded by a 48-hour interruption. EAM data were measured a median of 52 days (IQR: 32-78) and 84 days (IQR: 75-88) of the 90 days preceding a viral load test in the standard and real-time EAM datasets, respectively. During standard monitoring, during the 90 days preceding a viral load test median average adherence was 89% (IQR: 75%-96%) with a median of 2 treatment interruptions $\geq$24 hours (IQR:1-4). During real-time monitoring, during the 90 days preceding a viral load test median average adherence was 93% (IQR: 86%-97%) with a median of 3 interruptions of $>$24 hours (IQR:1-8).

Table 2.1: Baseline[a] characteristics of the study population

	Standard EAM (N=443)	Real-time EAM (N=485)
Woman	N = 307 (69.3%)	N = 345 (71.1%)
Age (years)	median 35 (IQR: 30-39)	median 33 (IQR: 27-40)
Follow-up time (years)	median 4.6 (IQR: 2.5-5.6)	median 2.2 (IQR: 1.4-2.5)
CD4+ T cell counts (cells/ mm^3 at ART initiation)	median 135 (IQR: 78-202)	median 200 (IQR: 111-317)
NNRTI at baseline	N = 440 (99.3%)	N = 447 (92.2%)
Efiravenz	N = 57 (13%)	N = 228 (51%)
Nevirapine	N = 383 (87%)	N = 219 (49%)
Plasma HIV RNA level (viral load) (copies/ml) at ART initiation	median 113,888 (IQR: 39,789-34,3272)	median 94,041 (IQR: 30,631-299,705)
Total days from ART initiation to baseline[a]	median 168 (IQR: 164-175)	median 112 (IQR: 107-120)[b]
Initiated ART within 120 days prior to first EAM monitoring	N = 72 (16%)	N = 187 (39%)
Viral load tests included as outcomes	5,922 tests[c]	2,834 tests[d]
Participants continuing monitoring from standard EAM	NA	N = 243

Among HIV-infected adults followed with electronic adherence monitoring using either standard or real-time devices following ART initiation in Uganda.
Missing for standard EAM Users: Baseline Viral Load (n=8; 1.8%), female (n = 5; 1.1%), age (n=5; 1.1%). Missing for real-time Users: Baseline Viral Load (n=12; 2.5%), Baseline CD4 (n = 5; 1%)
NNRTI; nonnucleoside reverse transcriptase inhibitor. IQR; Interquartile Range
[a]Baseline: first viral load test while using electronic adherence monitoring. [b]Estimate is among participants continuing after being monitored by standard EAM. [c]15 tests (0.2%) excluded due to a greater than 400-day lapse in testing. An additional 300 tests (5%) excluded from machine learning training set, but not from evaluation of performance, due to no adherence data in at least 180 days or missing predictor data. [d]389 tests (12%) excluded due to a greater than a 400-day lapse in testing; 98% of these occurred during the final three months of study follow-up. An additional 168 tests (6%) excluded from machine learning training set, but not from evaluation of performance, due to no adherence data in at least 180 days or missing predictor data

Contribution of EAM to machine-learning-based prediction of viremia

Super learning applied to standard EAM, clinical, and demographic data ("Full EAM" predictors) yielded a cvAUC for viremia of 0.77 (95% CI: 0.72, 0.81), non-significantly (p=0.08)

higher than the cvAUC of 0.70 (95% CI: 0.64, 0.76) achieved using clinical predictors alone (Figure 2.1, Table 2.2). Addition of standard EAM data without viral loads to the clinical predictor set resulted in a modest but non-significant (p=0.27) increase in the cvAUC (0.75; 95% CI: 0.69, 0.80). Additional inclusion of baseline viral load or removal of CD4+ T cell count from the predictor set had a minimal impact on cvAUC.

Super learning applied to clinical predictors alone achieved a cvAUC in the real-time EAM data set of 0.78 (95% CI: 0.72, 0.85). In contrast to the modest improvement in performance achieved with standard EAM data, addition of baseline viral load and real-time EAM predictors to the clinical predictors resulted in a significant (p=0.03) improvement in the cvAUC (0.88 95% CI: 0.83, 0.93). Addition of real-time EAM data alone to the clinical predictors resulted in moderate improvement in the cvAUC (p=0.06) (0.86 95% CI: 0.81, 0.91). Removal of time-varying CD4 count from the "EAM + Clinical" set had no impact on the cvAUC, suggesting little incremental gain in prediction of viremia from this measure (Table 2.2, Figure 2.2). Comparisons of alternative predictor sets based on net reclassification improvement were qualitatively similar for both standard and real-time EAM; for EAM + Clinical predictor set (used in the EAM-based testing strategy) the Hosmer-Lemeshow test supported adequate calibration (Supplement 2).

Potential performance of a EAM-guided differentiated monitoring strategy

We evaluated hypothetical rules for triggering viral load tests based on combining the machine learning risk score with a cutoff above which a viral load test would be ordered. The cutoff was chosen to meet the sensitivity-constrained criteria (Table 2.2, Figure 2.3). When trained on standard EAM data, the super learner attained a cross-validated sensitivity of 93-96% and rate of negative prediction of 25-31%. When trained on real-time EAM data, super learner attained a cross-validated sensitivity of 88-91% and rate of negative prediction of 37-47%.

Finally, we compared the performance of two non-differentiated strategies to a modified version of the machine-learning classification procedure described above (Table 2.3). When based on standard EAM data, the "3-Month" schedule would have reduced the number of tests ordered by 3% and would have delayed detection of 2% of observed viremia cases, for an average delay in detection among all viremia cases of one day. In contrast, the "WHO" schedule would have reduced the number of tests ordered by 77%, but would have resulted in the delayed detection of 67% of viremia cases, with an average delay in detection of 61 days. Finally, the "EAM" machine-learning approach could have reduced the total number of viral load tests ordered by 24% while delaying detection of only 9% of viremia cases, with an average delay in detection of 9 days. Under "3-month," "WHO," and "EAM" schedules, 97%, 22%, and 76% of non-viremic cases would have received a test, respectively.

Using the real-time EAM data, the "3-Month" schedule would have avoided 19% of observed tests (a reduction relative to the observed schedule due to eliminating extra tests triggered by detected interruptions) and delayed detection of 16% of viremia cases by 8 days on average, while the "WHO" schedule would have reduced the number of tests by 69%, and

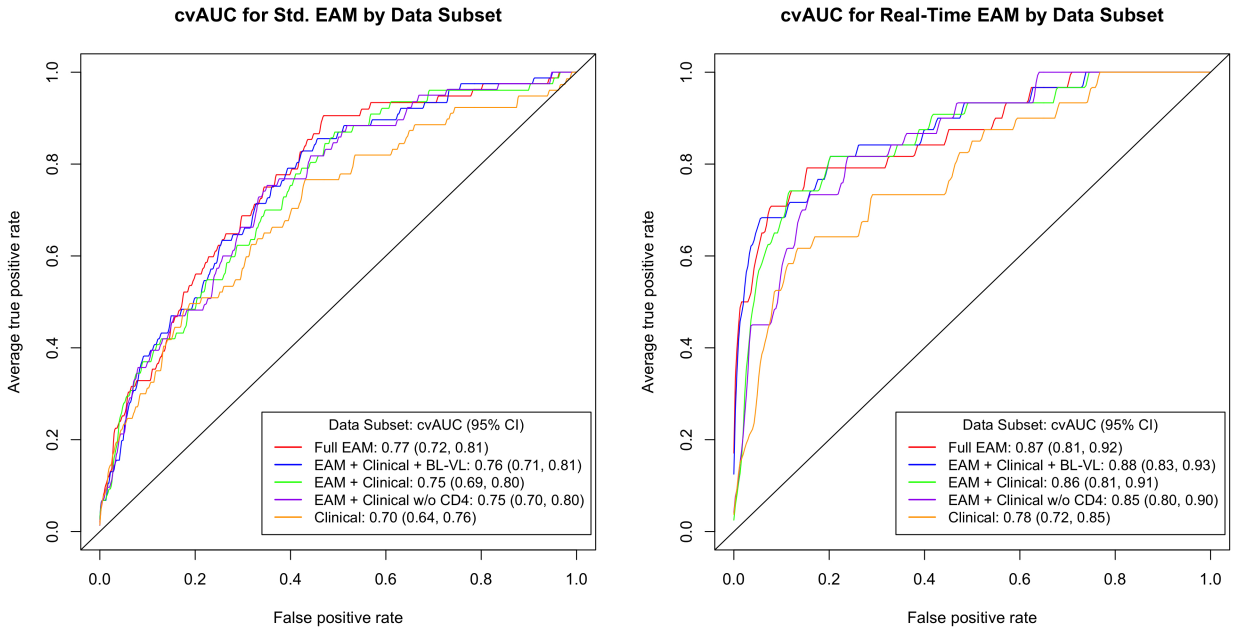


Figure 2.1: Cross-validated (10-fold) ROC curve for super learner prediction of viremia using 5 predictor sets (Supplement 1).

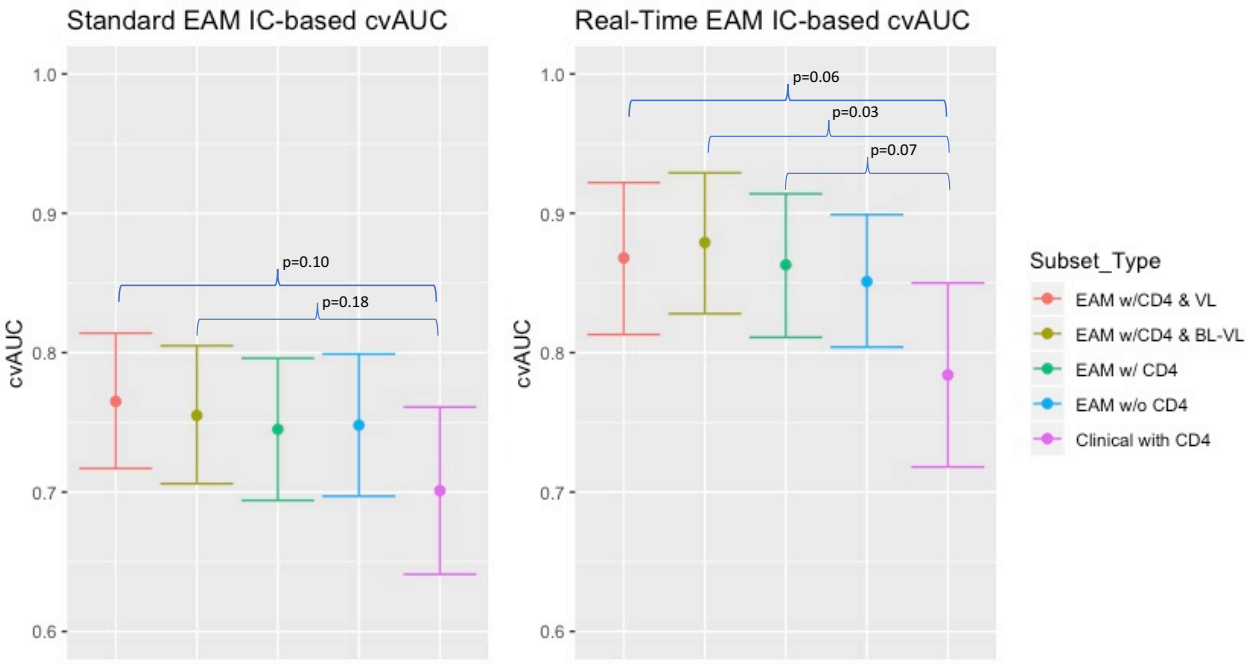
Viremia defined as HIV RNA level > 1000 copies per mL
blVL; baseline viral load
EAM; electronic adherence monitoring VL; viral load

delayed detection of 84% of viremia cases by 74 days on average. In contrast, the "EAM" approach would have avoided 32% of all viral load tests, while delaying detection of 13% of viremia cases with an average delay of 11 days. Under "3-month," "WHO," and "EAM" schedules, 81%, 32% and 68%, of non-viremic cases would have received a test, respectively.

2.4 Discussion

Analysis of real-time electronic adherence data using ensemble machine-learning achieved excellent prediction of viremia among HIV-infected individuals treated with ART in rural Uganda (cvAUC of 0.88). Addition of real-time EAM data together with viral loads to basic demographic and clinical data significantly improved prediction of viremia, indicating potential value added by this technology. However, further addition of post-ART CD4+ T cell counts to the predictor set did not significantly improve global predictive performance (as assessed with cvAUC), supporting prior findings of the limited value of CD4 count for predicting viremia [84–86].

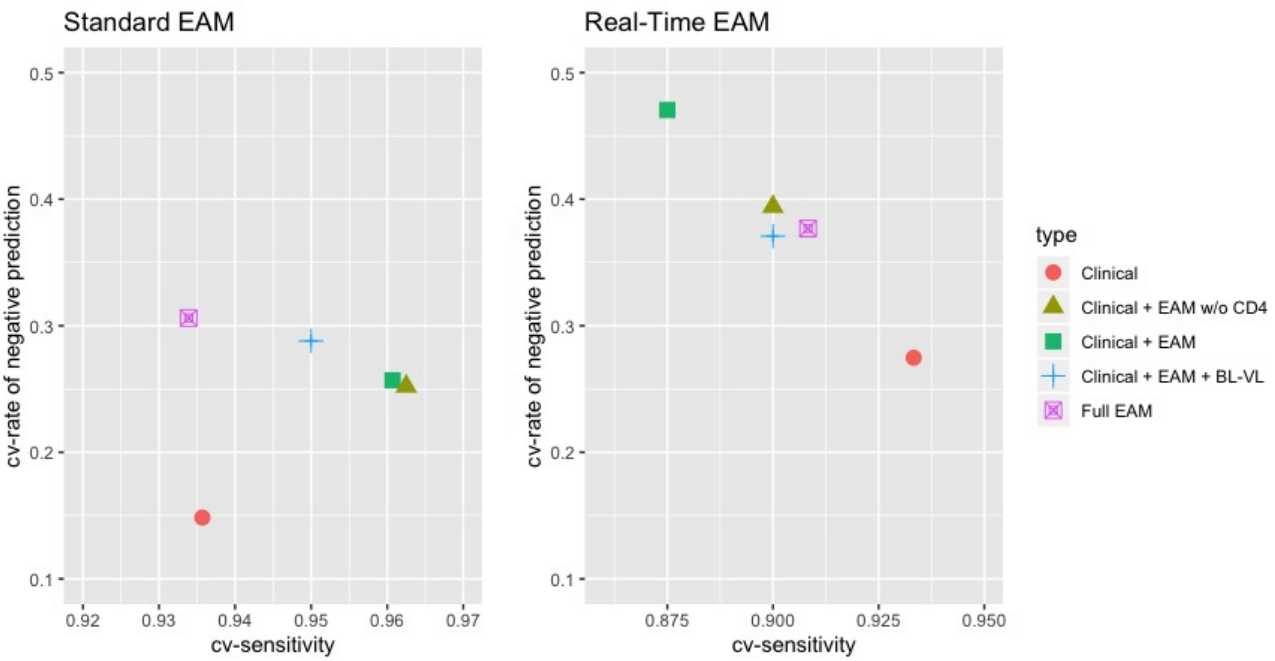
Figure 2.2: Influence-curve (IC) based inference; p-value and confidence interval based on variance of influence curve of AUC



Our results suggest that a testing strategy using real-time EAM to decide when to order versus defer viral load testing could substantially reduce the number of viral load tests ordered (32-47% of observed tests avoided, depending on the strategy and availability of viral load), while still detecting most viremia cases without additional delay. While some strategies incorporated baseline viral load, which may not be routinely available, the benefits of an EAM-based differentiated testing approach were also substantial when viral loads were not used. By comparison, the testing schedule recommended by the WHO for stable patients, if deployed uniformly for all participants in our sample, would have reduced the number of tests ordered by 69%, but would have delayed detection of viremia 74 days on average among individuals experiencing viremia. Extended viremia increases the risk of developing drug resistant virus [87], and of onwards transmission of HIV infection [88].

Super learning applied to standard EAM data, in combination with clinical and demographic data, was able to predict viremia well (cvAUC of 0.77), and would have avoided 25-31% of observed viral load tests while detecting most viremia cases without additional delay. In contrast, the improvement in prediction seen with addition of standard EAM data to clinical and demographic data was not statistically significant (p=0.08). Differences in the "value-added" of real-time versus standard EAM may have been due to differences in participant characteristics or temporal trends– participants in the real-time EAM cohort were

Figure 2.3: Classification: cv-sensitivity constrained rate of negative prediction



followed more recently and had higher CD4+ T cells at ART initiation. Implementation of real-time monitoring may also have provided better information to guide differentiated testing compared to standard monitoring by allowing for real-time data quality improvements (for example, identifying periods of device non-use). Further, interruptions in events could trigger additional tests during real-time but not standard monitoring; thus, both the reference "observed" testing regime differed and the extra tests themselves may have changed adherence. Indeed, average adherence appeared to increase when participants were switched from standard to real-time monitoring [27], and qualitative work supports a possible motivational effect of 'being watched' [48, 89]. However, the number of >24 hour interruptions was similar if not higher during real-time monitoring. These sustained interruptions may be related to structural barriers to adherence (e.g., lack of transportation to pick up medication) that are not as amenable to changes in motivation and instead reflect circumstantial differences in the two monitoring periods.

Our study has limitations. First, we assumed that excluding observed tests would not have changed subsequent adherence, viremia, or their relationship; to improve the plausibility of this assumption, we censored at first detected viremia. However, less frequent monitoring might affect adherence by reducing positive feedback provided to participants. Second, use of a real-time monitoring strategy to guide viral load testing would make possible not only targeted reduction in testing frequency, but also early adherence and testing interventions (beyond those implemented in the real-time EAM study); the current analysis

Table 2.2: Cross-validated AUROC, cross-validated rate of negative prediction, cross-validated sensitivity and cross-validated number needed to screen

	Full EAM	Clinical + EAM + + blVL	Clinical + EAM	Clinical + EAM, no CD4	Clinical
Std. EAM					
cvAUC	0.77	0.76	0.75	0.75	0.70
(95% CI)	(0.72, 0.81)	(0.71, 0.81)	(0.69, 0.80)	(0.70, 0.80)	(0.64, 0.76)
cvRNP	0.31	0.29	0.26	0.25	0.15
cvSens	0.93	0.95	0.96	0.96	0.94
cvNNS	51	52	54	55	62
cvFPR	0.69	0.71	0.74	0.75	0.85
cvPPV	0.02	0.02	0.02	0.02	0.02
Real-time EAM					
cvAUC	0.87	0.88	0.86	0.85	0.78
(95% CI)	(0.81, 0.92)	(0.83, 0.93)	(0.81, 0.91)	(0.80, 0.90)	(0.72, 0.85)
cvRNP	0.38	0.37	0.47	0.39	0.28
cvSens	0.91	0.9	0.88	0.9	0.93
cvNNS	48	47	41	46	56
cvFPR	0.62	0.63	0.53	0.60	0.72
cvPPV	0.02	0.02	0.02	0.02	0.02

EAM; Electronic Adherence Monitoring
VL; Viral load, blVL; Baseline Viral load
cvAUC; cross-validated area under ROC curve
cvRNP; cross-validated rate of negative prediction (proportion of tests avoided)
cvSens; cross-validated sensitivity
cvNNS; cross-validated number needed to screen
cvFPR; cross-validated false positive rate
cvPPV; cross-validated positive predictive value (precision)

is conservative in the sense that it does not incorporate these additional potential benefits. Third, while estimates of the proportion of viral load tests that could be deferred under hypothetical testing strategies were based on independent data (via cross-validation), accurate quantification of their uncertainty is an area of current work. Although few participants had missing viral load tests under the research protocol testing schedule, a future analysis should consider how missing outcomes affected results.

Finally, cost is an obvious concern when considering technology for clinical care. Cost-effectiveness is beyond the scope of this analysis; however, a cost-effectiveness analysis of potential ART adherence monitoring interventions in sub-Saharan Africa found that an adherence monitoring-based intervention could cost up to $50 per person-year on ART while re-

Table 2.3: Performance of "3-month," "WHO," and "EAM" testing rules

	Standard (EAM, CD4, w/o VL)				Real-time (EAM, CD4, w/o VL)			
	3-month	WHO	EAM[a]	Obs	3-month	WHO	EAM[a]	Obs
RNP	0.03	0.77	0.24	0	0.19	0.69	0.32	0
1-Sensitivity	0.02	0.67	0.09	0	0.16	0.84	0.13	0
No. Tests Total	5728	1333	4518	5922	2304	886	1928	2834
FPR	0.97	0.22	0.76	0	0.81	0.32	0.68	0
Mean, median delay time among undetected viremia cases, days (IQR)	42, 43 (37-48)	91, 96 (88-101)	97, 96 (89-97)	NA	49, 48 (31-64)	88, 76 (68-107)	84, 74 (73-75)	NA
Mean, median delay time among all viremia viremia cases, days (IQR)	1, 0 (0-0)	61, 88 (0-97)	9, 0 (0-0)	NA	8, 0 (0-0)	74, 73 (62-97)	11, 0 (0-0)	NA

3-Month; Test every 3 mos,
WHO; Test 6 and 12 mos after ART initiation and yearly thereafter
EAM; WHO schedule with additional testing if algorithm predicts high risk of viremia
Obs; Observed testing schedule for performance reference
RNP; rate of negative prediction
FPR; False positive rate
[a]EAM performance metrics differ slightly from Table 2.2 due to testing schedule augmented by "WHO" schedule (See Methods)

maining cost-effective, mainly driven by savings through effective differentiation of care [63]. Current lower cost versions of real-time adherence monitoring devices consistent with that threshold [90] are now available and are being tested for use in routine care (NCT03825952). Application of a previously-developed machine learning tool does not require intensive computing resources [91], and the increasing use of smart phones globally [92] could make even real-time updates to a machine learning algorithm feasible in the foreseeable future. Further, a differentiated care approach could be used to target use of these devices (e.g., in patients with self-reported adherence challenges).

2.5 Conclusions

Evidence is increasing that differentiated care for HIV patients in resource-limited settings is a cost-effective intervention [13, 63]. Real-time electronic adherence monitoring provides one possible tool to support a differentiated care strategy by making it possible to offer viral load testing and adherence interventions on an individualized schedule in response to evolving patient needs. Similar low-cost, real-time technology is being used in clinical care for tuberculosis and is being assessed for ART in Uganda [39]. This technology allows providers

to triage at-risk patients. However, the informative real-time data that rapidly accumulate through these devices may be difficult to interpret. Flexible algorithms with the capacity to leverage these data, such as those presented here, could be readily integrated into accessible software to address this issue.

In conclusion, our analysis suggests that real-time electronic adherence data analyzed with machine-learning methods have potential to achieve gains in the efficiency of a targeted viral load monitoring strategy while maintaining high sensitivity for detection of viremia. Future work should prioritize external validation of differentiated strategies in new settings, and implementation of these methods through software that aims to guide differentiated patient care. Our results provide an illustration of the utility of machine-learning methods to better leverage complex data for precision medicine and public health.

2.6 Supplement I

Supplemental Material for Super Learning Analysis of Real-Time Electronically Monitored Adherence

---

Training Sets

Various training sets were created by merging all or subsets of the following data sources: demographic, baseline clinical, time-varying clinical, time-varying viral load data, ART and electronic adherence monitoring (EAM) summaries.

The following distinct training sets were considered (subsets of the predictor variables):

(a) All predictors.

(b) Remove time-varying viral load data.

(c) Remove time-varying viral load data, and baseline viral load.

(d) Remove time-varying viral load data, baseline viral load, and time-varying CD4.

(e) Remove time-varying viral load data, baseline viral load, and EAM.

Data Descriptions

Demographic / Baseline CD4

- **Age**: Patient age (years).
- **Sex**: Binary indicator of female gender.
- **Baseline CD4**: Last measured CD4 with date $\leq$ ART start date.

Time-varying CD4

- **Last CD4**: Last measured CD4 value.
- **Time Since CD4 Measured**: Number of days since CD4 measurement was recorded.
- **Nadir CD4**: min of all CD4s with date $\leq$ current viral load (VL) date.

Baseline Viral Load

Let Q_{pre} be the last VL test date prior to the first regimen start date.

- **Days between first regimen start date and** Q_{pre}: First regimen start date - Q_{pre} (days)
- **Time since** Q_{pre}: Current (row) date - Q_{pre} (days)

Time-varying Viral Load

- **Last Viral Load**: Previous VL result.
- **Last Viral Load Test at Limit**: Binary indicator of previous VL result at the limit of detection.
- **Time Since Last Viral Load Test**: Number of days since last VL test.
- **Previous VL** $>$ **400**: Number of previous VL tests that were $>$ 400 copies/ml, which occurred prior to current visit and $\geq$ 90 days after ART start date.

ART

- **Time On Study**: Days since start of EAM (either standard or real-time).
- **ARV Code**, binary indicator for each of the following ARV codes (21 total*): 3TC, ABC, ALU, ATR, ATZ, AZT, AZR, D4T, DUN, DUO, EFV, LAS, LPV, NVP, RAL, RTV, TDF, TLA, TLE, TRI, TRU
- **Drug Class**, binary indicator for each of the following drug classes: NRTI, PI, NNRTI, and Other
- **ARV Regimen**, binary indicator columns, where regimen varies over the unique ARV code combinations (53 total).
- **Med Class** for $X \in \{0, 1, 2, 4\}$: Regimen class* (no instances of type 3 or 5 in training data). Note that this is 0 only prior to ART start. These are binary indicator columns, but each row belongs to only one med class.
- **Num Drugs Current**: Number of unique ARV codes in current regimen.
- **Num Drug Classes Current**: Number of unique drug classes in current regimen.
- **New Drug Introduced**: Binary indicator of a new ARV code (new = 1). All visits prior to and including ARV start date are coded as 0.

- **New Regimen Introduced**: Binary indicator of a new ARV code based regimen (new = 1). All visits prior to and including ARV start date are coded as 0.
- **Num Regimen Changes**: Number of changes in ARV code-based regimens since ART start (change defined as: current ARV code != Prior ARV codes)
- **Num Med Classes Ever**: Number of unique med classes since ART start.
- **Num Drug Classes Ever**: Number of unique drug classes since ART start.
- **Num Drugs Ever**: Number of unique ARV codes ever used.
- **Daily Total Doses**: Number (0-10) of doses prescribed per day across all medications. (bid=2, qd=1)
- **Dose Frequency**: Maximum number (0-2) of daily doses among the unique medications. (bid=2, qd=1)
- **Any X** for X $\in$ {3TC, TDF, AZT, D4T, EFV, NVP, FTC}.*
- **Baseline Year**: Calendar year at ART start date.

EAM (Adherence Summaries)

- **Days Since Last Monitoring**: Current date $-$ date of previous monitoring event.
- **Avg Adherence X** for $X \in \{7, 14, 30, 60, 90, 180, 365\}$: Average adherence over the last X days.
- **Avg Adherence VL**: Average adherence since the last VL test.
- **Days Monitored X** for $X \in \{7, 14, 30, 60, 90, 180, 365\}$: Number of days monitoring data exists over the last X days.
- **Min Adherence X, N** for $X \in \{7, 14, 30, 60, 90, 180, 365\}$ and $N \in \{2, 3, 4, 7, 14, 30\}$: Minimum adherence over the past X days using N-day sliding windows, such that $X > N$.
- **Min Adherence VL N** for $N \in \{2, 3, 4, 7, 14, 30\}$: Minimum adherence since the last VL test using N-day sliding windows.
- **Variance of Adherence X** for $X \in \{7, 14, 30, 60, 90, 180, 365\}$
- **Variance of Adherence VL**: Variance of adherence since the last VL test.
- **Num Interruptions X, Y** for $X \in \{7, 14, 30, 60, 90, 180, 365\}$ and $Y \in \{1, 2, 3, 4, 7, 10, 11, 12, ..., 19, 20, 25, 30\}$: Number of interruptions over the past X days, of at least Y days.

- **Num Interruptions VL Y** for $Y \in \{1, 2, 3, 4, 7, 10, 11, 12, ..., 19, 20, 25, 30\}$: Number of interruptions since the last VL test, of at least Y days.
- **Variance of inter-dosing interval in last X days** for $X \in \{7, 14, 30, 60, 90, 180, 365\}$: Variance of inter-dosing interval over the last X days.
- **Variance of Inter-Dosing intervaI**: Variance of inter-dosing interval since the last VL test.

Training Set: Predicting Viremia

The training data sets are constructed from the original viral load (VL) test result data. Each row corresponds to a clinic visit with a non-missing viral load result. Each patient has multiple visits, so we treat the data as a pooled repeated measures data set. Rows are uniquely determined by patient ID and visit date. If the outcome is missing, the sample will be excluded. Additionally, visits for which the EAM data is missing or outdated (more than 180 days old) are excluded.

Start Date

The start date is the first visit which is at least 90 days after ART start date, for which there is prior EAM data.

End Date

The end date is the earliest of: {date of first detected viremia, last visit}

Imputation

EAM variables were imputed using the most recent available value. For example, if `avgAdh_7` and `avgAdh_14` were missing, but `avgAdh_30` was available, then both would be imputed with the `avgAdh_30` value. Only EAM variables were imputed, and the remainder of the rows that still contained missing values in non-EAM columns were removed.

medClass	Description
0	Has not yet started treatment
1	Any NNRTI and no PI
2	Any PI and no NNRTI
3	Both NNRTI and PI
4	All NRTI
5	Other

Table 2.4: Definition of medClass variable

Variable	Associated ARV codes
any3TC	x3TC, DUO, TLE, TRI
anyTDF	TDF, ATR, TLE, TRU
anyAZT	AZT, DUO, DUN
anyD4T	D4T, TRI
anyEFV	EFV, ATR
anyNVP	NVP, DUN, TRI
anyFTC	TRU, ATR

Table 2.5: Definition of anyX set of variables

2.7 Supplement II

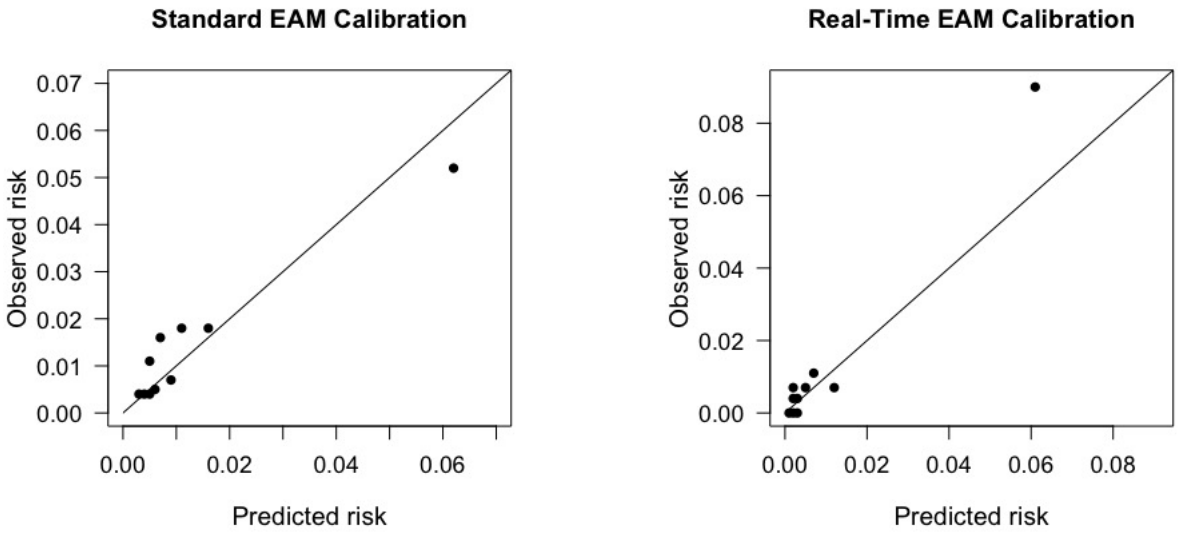
In addition to testing the differences in cvAUC between data subsets using the standard error estimates based on the cvAUC's influence function, the change in classification performance of the super learner based risk predictor performance across sets of predictor variables was also tested using the net reclassification improvement (NRI) statistic. The NRI assesses a prediction model's classification performance relative to a reference model by quantifying the new model' s changes in risk score assignment. Changes in risk classification occur when the prediction model being compared categorizes an individual to a higher or lower risk, relative to the reference model [93]. The NRI analysis led to results that were similar to those obtained when the change in cvAUC was tested using the influence function.

Table 2.6: Net reclassification improvement (NRI) for standard and real-time EAM prediction by data subset

	E-A	E-B	E-C	E-D
Std. EAM NRI	0.243	0.138	0.151	0.036
(p-value)	(0.034)	(0.226)	(0.186)	(0.754)
Real-time EAM NRI	0.750	0.921	0.634	0.523
(p-value)	(0.000)	(0.000)	(0.000)	(0.002)

Relative to Data Subset E (Clinical)
A; Full EAM
B; Clinical + EAM + blVL
C; Clinical + EAM
D; Clinical + EAM, no CD4

Figure 2.4: Super learner calibration plot



(a) Standard EAM

(b) Real-time EAM

Table 2.7: Hosmer-Lemeshow test for standard EAM

risk score decile	total obs.	mean pred.	mean obs.	pred. # cases	obs. # cases
[0.0016,0.0037)	563	0.003	0.004	1.81	2
[0.0037,0.0044)	562	0.004	0.004	2.30	2
[0.0044,0.0051)	562	0.005	0.004	2.66	2
[0.0051,0.0058)	562	0.005	0.011	3.05	6
[0.0058,0.0067)	562	0.006	0.005	3.52	3
[0.0062,0.0079)	563	0.007	0.016	4.10	9
[0.0079,0.0098)	562	0.009	0.007	4.91	4
[0.0098,0.0127)	562	0.011	0.018	6.26	10
[0.0127,0.0215)	562	0.016	0.018	9.04	10
[0.0215,0.6031]	562	0.062	0.052	34.92	29

Based on EAM, CD4, w/o VL
Using 10 risk score bins; Hosmer-Lemeshow goodness of fit p-value: 0.106
total obs.; total number of observations in risk decile
mean pred.; mean number of predicted viremic cases in risk decile
mean obs.; mean number of observed viremic cases in risk decile
pred. # cases; predicted number of viremic cases in risk decile
obs. # cases; observed number of viremic cases in risk decile

Table 2.8: Hosmer-Lemeshow test for real-time EAM

risk score decile	total obs.	mean pred.	mean obs.	pred. # cases	obs. # cases
[0.00093,0.0014)	267	0.001	0.000	0.33	0
[0.00142,0.0017)	267	0.002	0.004	0.42	1
[0.00171,0.0021)	266	0.002	0.000	0.50	0
[0.00205,0.0025)	267	0.002	0.007	0.60	2
[0.0025,0.0030)	266	0.003	0.000	0.72	0
[0.0030,0.0039)	267	0.003	0.004	0.90	1
[0.0039,0.0054)	267	0.005	0.007	1.22	2
[0.0054,0.0086)	266	0.007	0.011	1.81	3
[0.0086,0.0174)	267	0.012	0.007	3.20	2
[0.0174,0.6183]	266	0.061	0.090	16.24	24

Based on EAM, CD4, w/o VL
Using 10 risk score bins; Hosmer-Lemeshow goodness of fit p-value: 0.1758
total obs.; total number of observations in risk decile
mean pred.; mean number of predicted viremic cases in risk decile
mean obs.; mean number of observed viremic cases in risk decile
pred. # cases; predicted number of viremic cases in risk decile
obs. # cases; observed number of viremic cases in risk decile

Chapter 3

Prediction of gestational age in LMIC using ensemble learning

3.1 Introduction

Preterm birth is a leading determinant of neonatal death and death in children under five years [94]. Of the 15 million global preterm births which occur yearly, approximately 90% occur in low-middle income countries (LMIC) [34, 95]. Many infants who survive preterm birth are impaired with cognitive and physical disabilities [96, 97]. While interventions exist to prevent or help manage preterm birth complications, this requires accurate identification of preterm birth in the time leading up to the intrapartum period. After the intrapartum period, accurate reporting of hospital- and community-level preterm birth rates serves to inform resource allocation and policy in LMIC. Accurate gestational age (GA) measurement is thus essential both during pregnancy and after delivery. However, GA measurement is prone to several well-known limitations in developed countries and LMIC alike [98–101].

Several factors contribute to the difficulty of accurate GA measurement. Although ultrasound-based GA assessment in the first trimester is recognized as the "gold-standard," it becomes less accurate as gestation progresses [102]. As a result, women without access to or late entry into obstetric care will receive a less accurate assessment. Additionally, in resource-limited settings and LMIC, ultrasound may not be available due to high costs of equipment and of technician training [101], and women in LMIC are more likely to seek antenatal care later in pregnancy for several reasons [103]. When early ultrasound is unavailable, many clinicians estimate GA using the last menstrual period (LMP) or symphysis-fundal height, which each have their own set of limitations [101]. Postnatal assessments of the newborn, such as Dubowitz and Ballard, while informative, are less accurate in the case of intrauterine growth restriction [104].

Alternative dating methods have been explored in order to overcome the limitations described above. GA dating using routine newborn metabolic screening can predict risk of preterm birth with excellent sensitivity and specificity [98], and therefore can provide

high quality surveillance of preterm birth rates in communities. However, the metabolic screening approach relies on a heel-stick blood draw between 12 hours and 8 days after birth, and therefore it is not meant to be a point-of care diagnostic tool to guide clinical decision-making. Additionally, its high overhead costs make it difficult to implement widely in resource-limited settings. A lower cost GA dating alternative was presented in Rittenhouse et al. [34], as an approach which could provide clinical guidance at delivery as well as preterm birth rate reporting. Researchers in this study applied machine learning methods to data collected from a cohort of women in urban Zambia. The resulting algorithm combined a set of accessible maternal and infant characteristics (LMP, birth weight, twin delivery, maternal height, hypertension in labor, and HIV serostatus) to predict risk of preterm birth, and achieved high sensitivity and specificity. While this set of predictors is generally more accessible, this type of algorithm has not been replicated under more routine data systems in sub-Saharan Africa, which generally cannot measure maternal-infant characteristics with the same granularity and level of quality control. Additionally, this has not been replicated in rural and peri-urban settings in other regions of sub-Saharan Africa to account for underlying community differences which affect birth outcomes.

We leveraged ultrasound examination data from two Rwandan cohort studies to build a prediction algorithm that is similar to the Zambia algorithm. The two cohorts differed in level of ultrasound examination quality control and in size. Both cohorts consisted of women receiving ANC in health facilities across rural and peri-urban Rwanda. The first cohort consisted of women who enrolled in the Preterm Birth Initiative (PTBi), which was a cluster randomized trial conducted in participating health facilities across Rwanda. This study was designed to evaluate the impact of a group antenatal care (ANC) on GA at birth compared to mothers exposed to standard one-on-one ANC. Secondly, we considered a smaller centralized cohort study which aimed to determine the association between infections, micronutrients, and preterm birth in peri-urban Rwanda. In the context of these two studies and more broadly, accurate GA measurement plays an essential role for guiding clinical care surrounding delivery, and for informing accurate study results, future interventions, and policy.

We applied machine learning methods to both cohort datasets in parallel to develop a low-cost algorithm with the potential improve the current dichotomous infant classification of preterm or full term, and the potential to be applied to non-ultrasound sites in the future. In Section 2 of this manuscript, we describe the study populations of both the PTBi and 'peri-urban' study and the statistical methods used to build and evaluate the prediction algorithm. In Section 3, we present the prediction performance of the algorithm for each cohort, based on different subsets of maternal-infant characteristics. Lastly, we conclude with a discussion in Section 4.

3.2 Methods

PTBi Study

Setting

We analyzed data from 18 PTBi study sites administering ultrasound examinations in five municipal districts across Rwanda between May 2017 and December 2018 (NCT03154177). These types of health facilities are the first point of contact for women seeking medical services, as they are generally easier for women to access compared to their district hospital. At the health facilities, nurses and midwives offer universal access to antenatal and postnatal care [50]. The health facilities varied in terms of rural or urban setting, monthly ANC volume, delivery volume, staff to patient ratio, and baseline PTB rate. Approximately three ANC providers per health facility were trained to use and conduct ultrasound, and additional providers were trained when possible to ensure the continuity of the service [50].

Population

Women seeking ANC at these facilities were eligible to enroll in the study if they attended their first ANC visit before 24 weeks gestation (as assessed by a clinician using LMP). To be included in this analysis, a mother must have received an ultrasound examination before 16 weeks GA (based on adjusted ultrasound GA at the time of examination). Additional exclusions were made if a maternal-infant record was missing a valid enrollment, ultrasound, or delivery date. Because general protocol at these health facilities is to refer cases of multiple gestation to the district hospital, we only included singleton pregnancies in this analysis. A maternal-infant record was excluded if GA at delivery was not between 24 and 42 weeks. Records lacking a complete set of maternal and infant characteristics were excluded, as these were required for building our algorithm. If fewer than 11 ultrasounds were recorded at any participating health facility, we did not include records from these facilities for the purposes of training the algorithm, resulting in exclusion of data from five health centers.

Peri-urban Cohort

Setting

Between September and October 2017, data collection for a separate study in the peri-urban Gasabo district of Kigali, Rwanda was ongoing. This smaller cohort study was designed independently of the PTBi study to investigate associations between maternal genitourinary infections, micronutrient deficiencies, and risk for preterm birth [49]. The study recruited women seeking ANC in 10 health centers in the Gasabo District. Participants were provided with transportation to and from a central study site, where assessments were performed by two obstetricians and two midwives in a supervised and standardized manner.

Population

Singleton pregnancy and gestational age between 9 to 20 weeks were confirmed by ultrasound and LMP for each woman who enrolled in the study. Maternal-infant records required a complete set of maternal and infant characteristics to be included. For inclusion in the GA prediction analysis, we only included records for women whose ultrasound examination was completed before 16 weeks gestation.

Observed Data

For both Rwandan cohorts, the observed data consisted of maternal and infant characteristics, where the independent unit of observation is the mother-infant dyad. For each mother attending her first ANC visit, her individual-level covariates were collected. Then, for all deliveries listed in maternity registers, infant characteristics and outcomes were captured. We considered the following maternal-infant characteristics to assess their value in predicting the outcome of interest, GA at delivery as measured by ultrasound. We focused on the subset of characteristics which overlapped between the two cohorts and which are known to be predictive of preterm birth, more specifically defined below.

Predictors

1. **Maternal characteristics** included: mother's age category, mother's stature < 150cm, and mother's mid-upper arm circumference (MUAC) < 21 cm. We considered the following self-reported socioeconomic status predictors: education level, occupation category, and indicator of alcohol use. We included parity and other obstetric history, including previous preterm delivery and previous stillbirth.

2. **Infant characteristics** measured at delivery: infant length, infant weight, sex, APGAR score (1 min), and month of delivery

Two additional characteristics were available in the PTBi cohort to predict ultrasound-based GA at delivery. The first was the mother's menstrual period (LMP). Secondly, GA at delivery was assigned by the birth care provider and recorded in the facility's maternity register. We will refer to this as 'recorded' GA throughout. Birth care providers used all available data to estimate this GA (LMP, birth weight, observed infant maturity), but GA assignment was not standardized or monitored across facilities [50]. In the the peri-urban cohort in Rwanda, 'recorded' GA was not collected in this study. LMP data for the peri-urban cohort is forthcoming.

Outcomes

We considered ultrasound-based GA at delivery as both a continuous and binary outcome: (i) infant's ultrasound-based GA at delivery in weeks, and (ii) binary indicator for whether infant's ultrasound-based GA at delivery was less than 37 weeks.

In the PTBi study, the outcome of interest was obtained without cross-checking first-trimester ultrasound with LMP. Although LMP-based GA is often combined with ultrasound-based GA depending on their agreement [100], we defined ultrasound-based GA independently of LMP-based GA, in order to assess the added value of incorporating LMP in a prediction algorithm.

In the peri-urban cohort, GA at delivery *was* defined using first trimester ultrasound cross-checked with reported LMP. Therefore LMP could not be used as an independent predictor.

Statistical Methods

INTERGROWTH-21st Standards

We assessed outliers in our data based on the INTERGROWTH-21st birth weight standards, which provided international anthropometric standards to better guide clinical assessments of infants across multiethnic populations and to complement current WHO standards [105]. The INTERGROWTH-21st project calculated sex-specific centiles for weight, length, and head circumference given gestational age at birth, based on eight geographically defined urban populations. The women in these locations shared similar maternal health and nutrition characteristics and had access to adequate ANC [105]. While we anticipate some differences in the PTBi and peri-urban cohort relative to the INTERGROWTH standards (for example, the INTERGROWTH sites were strictly urban), the guidelines served as a reference for assessing the likelihood of a data entry error or of an outlier in our population. Therefore, to better understand differences in our study populations, we estimated the percent of infants in each cohort that fell within the INTERGROWTH standards.

Predictor Subsets

We estimated the conditional expectation of the infant's ultrasound-based GA at delivery, and the infant's conditional risk of preterm birth, using the continuous and binary outcomes, respectively. To better understand which variables in our data are most predictive of ultrasound-based gestational age and preterm birth, we assessed the several subsets of the data. Using the PTBi data, we considered the following predictor sets (i) infant and maternal characteristics, (ii) LMP-based GA and maternal characteristics, (iii) LMP-based GA only, (iv) 'recorded' GA estimate only, (v) maternal characteristics only, and (vi) infant characteristics only. Each of these models adjusted for health facility to account for clustering effect.

The covariates collected from the peri-urban cohort differed slightly relative to the PTBi cohort, as described above. Therefore, the prediction models we considered were different than the PTBi models. For the peri-urban cohort, we considered the following covariate sets (i) infant and maternal characteristics, (ii) infant and maternal excluding infant length

(iii) infant and maternal excluding infant birthweight, (iv) infant and maternal excluding APGAR score, (v) maternal characteristics only, and (vi) infant characteristics only.

Prediction using super learner

We trained an ensemble of algorithms to predict ultrasound-based GA for each cohort of women described. We used ensemble learning, specifically, super learner, to build the prediction algorithm. Super learner uses internal cross-validation to combine an ensemble of machine learning algorithms, with the objective of by minimizing the cross-validated risk corresponding to a user-selected loss function [21]. For a continuous outcome, we used a non-negative least squares loss function. For a binary outcome, we used the negative log likelihood loss. The loss function can also be tailored based on the research question, for example, maximizing the sensitivity-constrained specificity or the positive predictive value. These will be considered in sensitivity analyses [technical Appendix].

The algorithms considered for GA prediction were: the elastic net generalized linear model [74], Bayes generalized linear model [73], random forest [72], polynomial spline regression [106], and partitioning and regression trees [107].

Performance metrics

To evaluate the performance of the prediction model (built using internal cross-validation), an additional layer of external cross-validation was used. The full super learner algorithm was run 10 times on 9/10th of the data in turn, and performance was evaluated by calculating performance metrics in each corresponding validation set and averaging. Using this approach, we plotted the cross-validated receiver operating characteristic (ROC) curve and estimated the cross-validated area under the ROC curve (cvAUC). We also plotted the precision-recall curve and estimated the cross-validated area under the precision-recall curve (cvAUPRC). Inference for this quantity based on the influence function is an area of ongoing research.

To assess the algorithm's capacity to classify preterm births at a specific risk-score cutoff, we chose the cutoff which yielded 80% specificity in each training set (a specificity threshold based on clinical standards). We estimated the performance metric in each validation fold using the corresponding training set's 80% specificity cutoff. Once this was done for each validation fold, we averaged across folds to attained the cross-validated performance metric. We estimated the following cross-validated metrics: the empirical sensitivity (# preterm infants classified as preterm/ # preterm infants) (cvSens), the empirical precision (# preterm infants classified as preterm/# infants classified as preterm) (cvPPV), and the false positive rate (#term infants classified as preterm/# term infants) (cvFPR).

To better understand which variables in our data are most predictive of gestational age and classification of preterm birth, we assessed the performance of the data subsets described of the metrics described above. Thus for each cohort, we assessed the performance of six different data subsets. Lastly, to assess the algorithm's performance in terms of predicting continuous GA in weeks based on the non-negative least square loss function, we plotted

the cross-validated predictions against the actual ultrasound-based GA. We reported the square-root of the mean squared error, to yield an estimate of the predictions' average error in weeks across all observations.

Ethics

In the PTBi study, women ages 15 and older who wished to enroll provided written informed consent between May 2017 and December 2018. Ethical approval was granted by the Rwanda National Ethics Committee (No.0034/RNE/2017) and University of California, San Francisco Institutional Review Board (16-21177). In the peri-urban cohort, the Institution Review Board (IRB) at the College of Medicine and Health Sciences (University of Rwanda) and the Research Unit at the Rwanda Ministry of Health (Approval notice: No213/ CMHS IRB/2017) approved the study protocol. All participants signed consent prior to participation [49].

3.3 Results

A total of 11,269 women sought ANC at PTBi facilities administering ultrasound during the study period. Of these women, 5,182 eligible women enrolled and attended ANC at the study sites, and consented to collection and inclusion of data in the analysis [50]. Of 795 women who received ultrasound in the first trimester, 185 (23%) were excluded from the sample due to missing data. This resulted in 610 maternal-infant records could be used to train the algorithm. The peri-urban study recruited 421 women seeking ANC, and 367 completed follow-up and met all inclusion criteria. Of the 296 women who received first trimester ultrasound, 8 (3%) were excluded due to missing data. This resulted in 288 maternal-infant records which could be used for the algorithm.

In the PTBi cohort, 85% of the newborn birthweights fell within the INTERGROWTH-21st 3rd and 97th percentiles, given the newborn's gestational age at delivery by first trimester ultrasound. In the peri-urban Rwanda cohort, 95.6% of newborn birthweights fell within the INTERGROWTH-21ST standards given the newborn's gestational age at delivery by first trimester ultrasound.

In the PTBi cohort, 6.1% of infants were born before 37 weeks GA, estimated by first trimester ultrasound alone. In the peri-urban cohort, 8.8% of infants were born before 37 weeks GA, estimated by first trimester ultrasound combined with LMP.

Table 3.1 shows the characteristics of each study population. The distribution across age categories was similar in both cohorts. 23% and 30% of women attended secondary school in the PTBi cohort and the peri-urban cohort, respectively. In the PTBi cohort, 6.1% of women had stunted growth, and 2.5% of women had MUAC of less than 21cm. In the PTBi cohort, 1.6% of women had history of preterm birth, and 3.1% had history of stillbirth. In the peri-urban cohort, 13.5% of women had stunted growth, and 1.1% of women had MUAC of less than 21cm. In the peri-urban cohort, 17.3% of women had history of preterm birth

and 3.5% had history of still birth. The two study populations were similar across all infant characteristics, including proportion of low birth weight (PTBi 7.2% and peri-urban 9.0%).

Figure 3.1 shows the global performance of the algorithm across multiple risk score cutoffs, using both the ROC and precision-recall curves. The performance was assessed using both the PTBi and peri-urban cohort data. The PTBi data subsets consisting of 'LMP and maternal' and 'LMP only' both achieved the highest cvAUC (cvAUC = 92%). The PTBi dataset including 'LMP + Maternal only' acheived the highest cvAUPRC (cvAUPRC=61%). The best performing peri-urban data subset in terms of both cvAUC and cvAUPRC was the subset using infant characteristics only (cvAUC =98%; cvAUPRC=84%)).

Table 3.2 shows the classification performance at a fixed threshold which achieves specificity of 80% in each training set. Using this approach, the algorithm based on the PTBi 'LMP only' data subset attained a cvSens of 87% and a cvPPV of 21%. Exclusion of the LMP and recorded GA (Infant and Maternal) greatly reduced the performance of the PTBi-based algorithm (cvSens=0.03; cvPPV=0.004). When based on the peri-urban cohort data, using the fixed specificity threshold approach, the algorithm attained a cvSens of up to 100%, and a cvPPV of up to 38%. Exclusion of infant characteristics (Maternal only) reduced the performance of the algorithm when trained on the peri-urban cohort data (cvSens=0.17; cvPPV=0.17).

Figure 3.2 and Figure 3.3 show the square root of the mean squared error of the GA predictions in weeks relative to the true ultrasound-based GA. When trained using the non-negative least squares loss function, the super learner's cross-validated predictions based only on the PTBi cohort's recorded GA resulted in an average error of 1.80 weeks across all predictions. When augmented with all available PTBi predictors, the super learner algorithm predictions resulted in an average error of 1.38 weeks, and 7% of the super learner's cross-validated gestational age predictions had a margin of error greater than 2 weeks. When trained on the peri-urban cohort data subset which excluded birth weight and infant length, the super learner predictions resulted in an average error of 1.96 weeks. When trained on the peri-urban set using all maternal-infant predictors, the super learner cross-validated predictions reduced the average error to 1.39 weeks, and 11% of the super learner's cross-validated gestational age predictions had a margin of error greater than 2 weeks.

3.4 Discussion

Our study sought to understand whether a simple subset of maternal-infant characteristics could be useful for predicting ultrasound-based GA at delivery in rural and peri-urban communities in Rwanda. Although many GA measurement techniques exist, they are prone to many limitations and may be particularly difficult to implement widely in LMIC. Further, accurate GA measurement is essential for reporting preterm birth rates in communities to help inform policy and resource allocation, and our algorithm may serve as an alternative method to assess GA.

Table 3.1: Characteristics of the study populations

	PTBi (N=610)	Peri-urban (N=288)
Maternal characteristics	n (%)	n (%)
Age < 20 years	40 (7.9%)	14 (4.9%)
Age 20-35 years	480 (78.7%)	237 (82.3%)
Age > 35	82 (13.4%)	37 (12.8%)
Attended secondary school	140 (23%)	87 (30.2%)
Alcohol use	116 (19%)	66 (22.9%)
Stature < 150 cm	37 (6.1%)	39 (13.5%)
MUAC < 21 cm	15 (2.5%)	3 (1.1%)
Parity	median 1 (IQR: 0-2)	median 2 (IQR: 1-2)
History of preterm birth	8 (1.3%)	49 (17.3%)
History of stillbirth	19 (3.1%)	10 (3.5%)
Infant characteristics		
Male	317 (52%)	135 (46.9%)
Length	median 49 (IQR: 48-50)	median 50 (IQR: 49-51)
Weight (kg)	median 3.1 (IQR: 2.9-3.5)	median 3.2 (IQR: 2.95-3.5)
Weight ≤ 2.5 kg	44 (7.2%)	26 (9%)
APGAR 1 min < 8	27 (4.4%)	16 (5.6%)

In the PTBi cohort, 185 of 795 eligible women were excluded due to missing values in covariates or excluded health facilities.
In the peri-urban cohort, 8 of 296 eligible women were excluded due to missing values in covariates.
MUAC; mid-upper arm circumference
APGAR; Appearance, Pulse, Grimace, Activity, and Respiration test score (scale 1-10)

We analyzed two different cohorts which consisted of women and their newborn infants across municipal districts in Rwanda. Using these data, we trained an algorithm and assessed its performance for prediction of GA among women who received first-trimester ultrasound. For each cohort, we considered sets of predictors which can be collected before delivery and others which can only be collected at delivery.

The algorithm based on the PTBi cohort data performed well when LMP was included in the predictor set in terms of cvAUC and cvAUPRC, for women who received first trimester ultrasound. Upon exclusion of LMP, performance dropped; infant and maternal characteristics alone performed poorly. In contrast, the algorithm based on the peri-urban cohort data performed well in terms of both of cvAUC and cvAUPRC when infant and maternal characteristics were both included among women who received first-trimester ultrasound. Inclusion

Table 3.2: Cross-validated metrics at fixed specificity (80%) by data subset

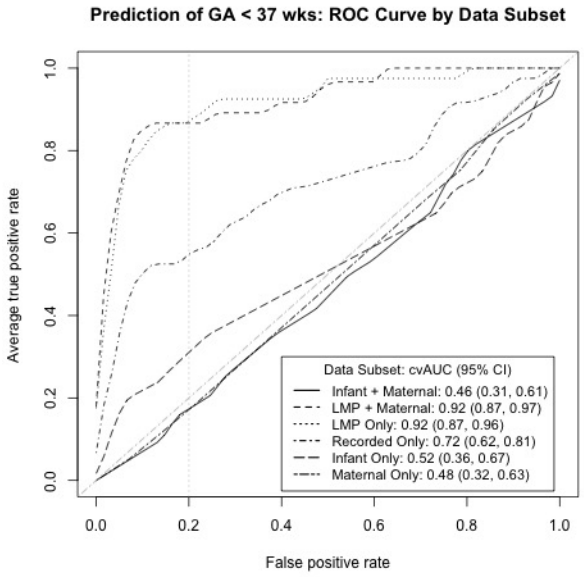
PTBi	cvAUC	CI.low	CI.high	cvSens	cvPPV	cvFPR
Infant + Maternal	0.462	0.314	0.611	0.033	0.004	0.061
LMP + Maternal	0.920	0.874	0.966	0.867	0.217	0.199
LMP Only	0.916	0.867	0.965	0.867	0.213	0.207
Recorded Only	0.716	0.621	0.810	0.525	0.152	0.200
Maternal Only	0.476	0.323	0.629	0.000	0.000	0.046
Infant Only	0.517	0.362	0.672	0.308	0.037	0.256
Peri-urban						
Infant + Maternal	0.970	0.951	0.989	1.000	0.331	0.202
No Length	0.940	0.900	0.979	0.883	0.346	0.201
No BW	0.954	0.928	0.979	1.000	0.378	0.202
No APGAR	0.963	0.940	0.986	1.000	0.337	0.201
Maternal Only	0.480	0.361	0.599	0.167	0.171	0.206
Infant Only	0.978	0.962	0.995	1.000	0.347	0.201

BW; birth weight
cvAUC; cross-validated area under receiver operating characteristic curve
cvSens; cross-validated sensitivity
cvPPV; cross-validated positive predictive value (precision)
cvFPR; cross-validated false positive rate

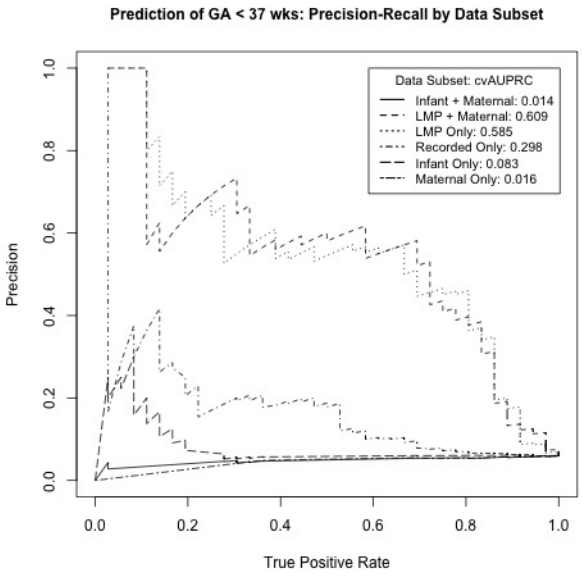
of LMP is forthcoming. However, upon exclusion of infant anthropometrics, the prediction algorithm performance significantly dropped, indicating the algorithm would not be useful for guiding clinical decisions but for improving preterm estimates. Lastly, because the results for continuous prediction of GA in weeks are within a two-week margin of error, then the algorithm may perform at least as well as other current clinical GA measurements, although this must first be validated in PTBi's non-ultrasound sites through a transportability analysis.

Interestingly, the results from both cohorts indicated that the models using larger sets of covariates did not necessarily add accuracy in cross-validated prediction of preterm births, contrary to the discussion in [34], which may indicate that, in our data, additional predictors added noise or caused overfitting (Table 3.2). Instead, the strongest signal appeared to come from LMP alone in the PTBi algorithm, and from infant anthropometrics in the peri-urban algorithm. LMP is often limited by poor recall and irregular cycle [100]. However, because this analysis was restricted to women who entered ANC at or before 16 weeks (who received first trimester ultrasound), then this cohort of women may be more highly motivated health consumers, who were able to report their LMP more accurately. It is possible that LMP

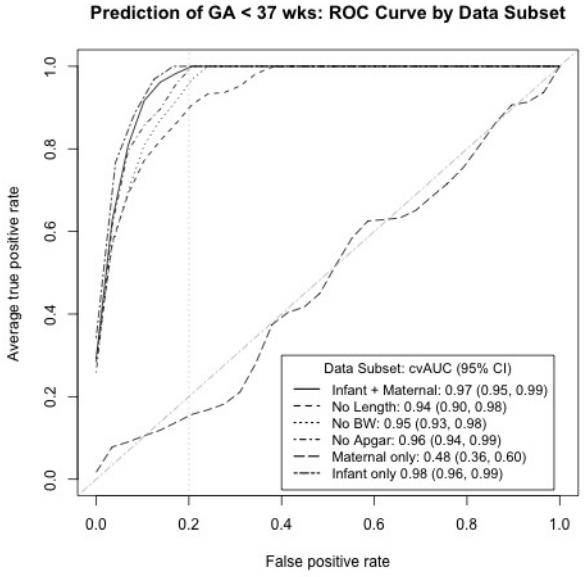
Figure 3.1: Binary prediction of preterm < 37 weeks GA at delivery



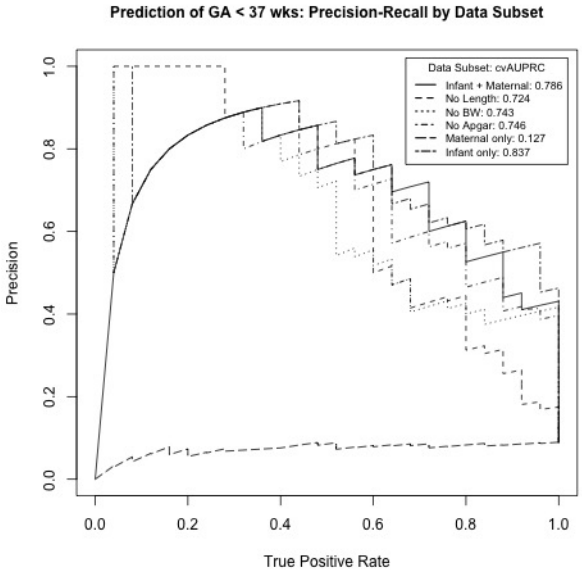
(a) PTBi ROC (N=610)



(b) PTBi Precision Recall (N=610)



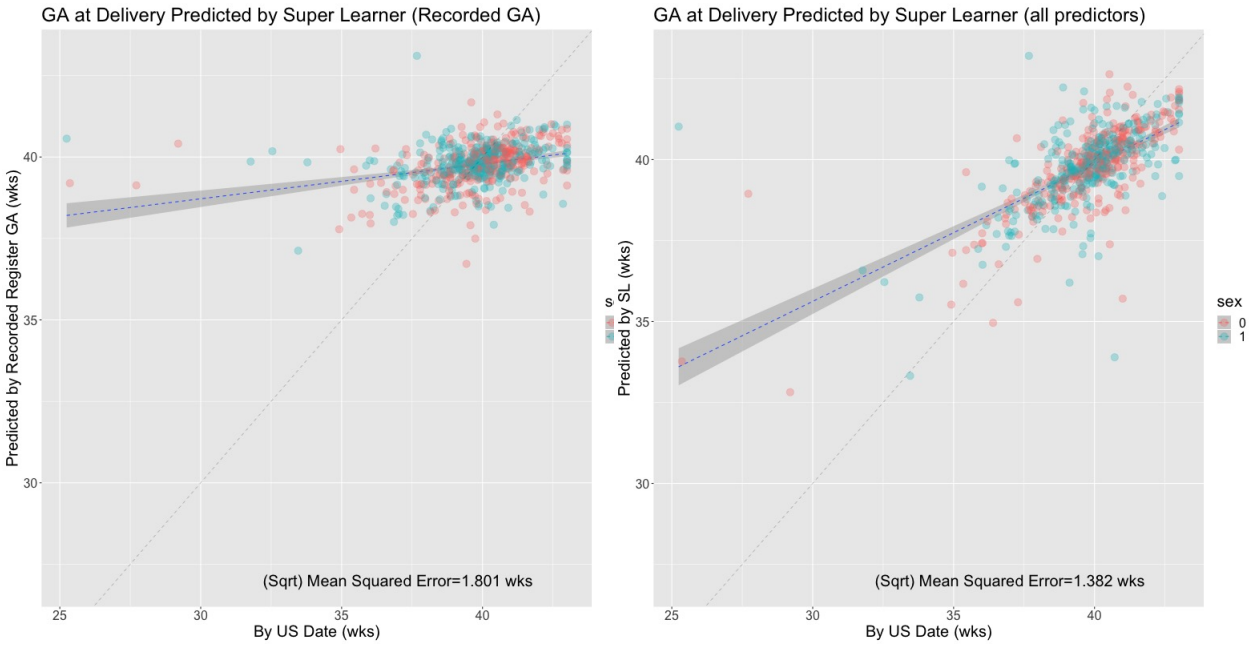
(c) Peri-urban ROC (N=288)



(d) Peri-urban Precision Recall (N=288)

would not perform as well among women who present for ANC later in pregnancy, and future analyses should be expanded to account for LMP and ultrasound recorded after the

Figure 3.2: Predicted GA at delivery in weeks for PTBi cohort



(a) PTBi GA weeks recorded only (N=610)

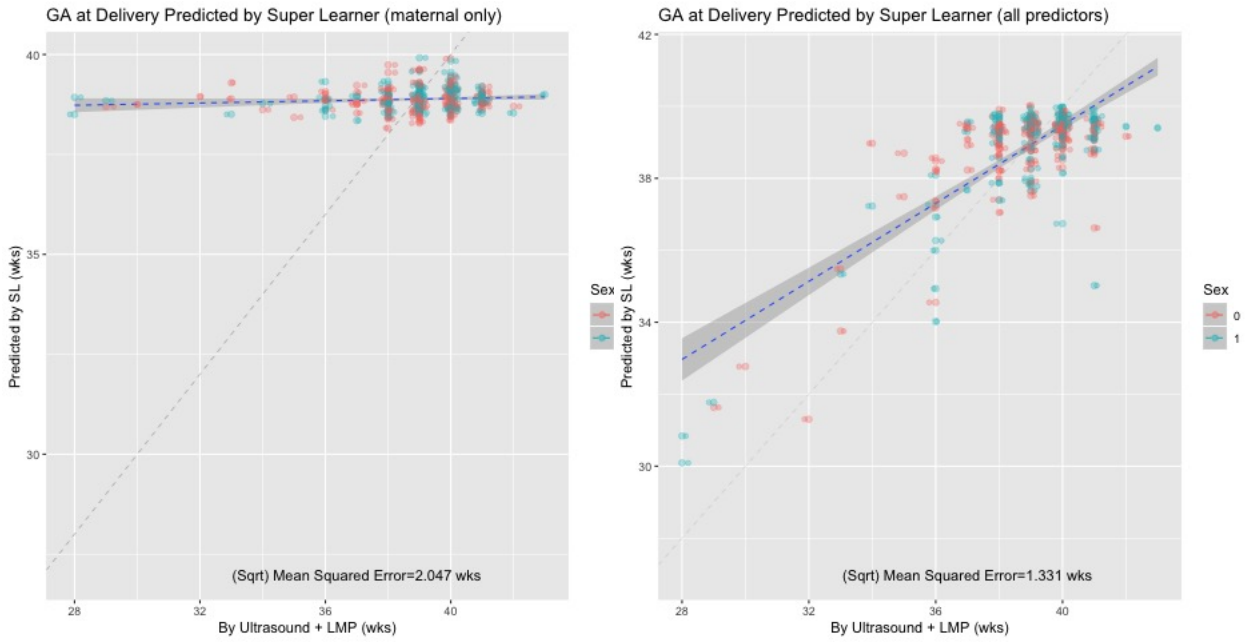
(b) PTBi All predictors (N=610)

first trimester. Raw LMP- and ultrasound- based GA data for the peri-urban cohort are forthcoming, since initially only the combined LMP and ultrasound GA at delivery was available as the primary outcome, and this will strengthen the comparisons we can make between the cohorts.

The algorithm based on the peri-urban cohort is likely to have performed better for several reasons. The study was smaller in scale, over a shorter period of time, and designed only for the purpose of obtaining gold-standard ultrasound to understand micronutrients and infections as they relate to GA. The PTBi study was much larger in scale and emphasis was designated towards primary outcomes that are related but adjacent to gold-standard ultrasound; the ultrasound data were intended for a secondary analysis. Further, clinical assessments of the women in the peri-urban cohort and subsequent data entry took place in a single location, and were carried out by a small team. Participants were provided transport to and from the centralized study site. The PTBi ultrasound examinations took place across 13 facilities over a longer period of time, and assessments were administered by several nurses and midwives who were also responsible for data entry. Thus, in the PTBi study, higher variability and lack of standardization in data entry were expected.

There are several limitations in this analysis. The INTERGROWTH standards may not be generalizable to rural communities in the PTBi study. However, these international guidelines helped us understand how the peri-urban and PTBi cohorts may differ. We emphasize

Figure 3.3: Predicted GA at delivery in weeks for peri-urban cohort



(a) Peri-urban No BW or Length (N=288)

(b) Peri-urban All Predictors (N=288)

that we cannot make inferences surrounding how the cohorts compare without a more rigorous analysis. The intention of analyzing these data in parallel is to better understand prediction of GA under two settings in Rwanda, and future work should investigate how to leverage the peri-urban cohort algorithm across different settings. In the PTBi study, ultrasound could not be performed at every site which may limit our understanding of the performance of the algorithm across all rural and urban communities in this study. Despite the potential systematic error in PTBi ultrasound measurement (based on low-level cadre, brief practical training, and pragmatic implementation) [52], the algorithm based on either LMP or recorded GA performed well. This indicates that, in the PTBi data, no other standard maternal-infant characteristic provided signal for predicting GA, especially relative to the peri-urban cohort. We hypothesize this may be due to potential data-entry errors, biases in data entry, or faulty connectivity at some facilities when uploading data [52]. While the distribution of PTBi maternal-infant characteristics does not appear to differ vastly compared to the peri-urban cohort (Table 3.1), a deeper data quality assessment is necessary.

Machine learning methods combined with data from the two cohorts demonstrated good overall performance in predicting preterm birth and gestational age at delivery in weeks, among women who received first-trimester ultrasound. This implies that the prediction algorithm could serve for multiple purposes, both monitoring and reporting, and future analyses should assess transportability of the algorithm beyond the PTBi-Rwanda non-

ultrasound sites. Based on a simple and parsimonious set of clinical predictors, the prediction algorithm may have the ability in the future to guide clinical decisions, validate database quality, and improve preterm birth estimates in settings where gold-standard ultrasound is not available. At a fixed specificity of 80% (20% FPR), the algorithm achieved high sensitivity. This threshold can be adjusted according to the desired use of the algorithm (e.g., routine clinical care, preterm birth rate reporting). For example, if this algorithm had been used to guide clinical care, resources such as medication and referral costs would have been erroneously allocated towards 20% of term infants. If users of the algorithm wish to minimize administering unnecessary treatments and costly referrals, future analyses should assess the performance of the algorithm at difference specificity thresholds. For potential future use by medical providers, the algorithm could be incorporated in software for use on a cell phone or tablet application. In LMIC, a GA algorithm such as the one that is presented here has the flexibility to be tailored for a specific population, and may avoid high costs and training needed for GA measurements that perform similarly.

Chapter 4

Comparative methods for cluster randomized trials

4.1 Introduction

Cluster randomized trials (CRTs) provide an opportunity to assess the population-level effects of interventions, which are randomly assigned to groups of individuals, such as communities, clinics, or schools. These groups are commonly called "clusters." In CRTs, the primary outcome may be defined at the individual-level or at the cluster-level, often as an aggregate of individual-level outcomes. The choice to randomize clusters, instead of individuals, is often driven by the type of intervention as well as practical concerns [108]. For example, interventions to improve medical practices are often randomized at the hospital or clinic-level to reduce logistical burden and to minimize the contamination between arms if individual patients were instead randomized. The design and conduct of CRTs has improved considerably [109–111], and results from CRTs have been widely published in public health, education, policy, and economics literature [112]. However, a recent review found that only 50% of CRTs were analyzed with appropriate methods [113].

The analysis of CRTs is complicated by the correlation of participant outcomes within clusters [108]. In other words, the cluster is the independent unit, and all observations within the cluster are dependent to some degree, which can present a challenge to statistical analysis and inference. Specifically, ignoring the correlation of individuals within clusters can lead to underestimates of the standard error. Ignoring the clustering may then magnify an intervention effect that is in fact attributed to the shared characteristics of a cluster, underestimating variance and inflating type I error. Clustering must be accounted for. Several approaches are available for this purpose; for example, using either a correction factor, or an unadjusted comparison of cluster-level outcomes, commonly implemented as the t-test, are just two simple approaches.

Once we have accounted for the clustering, the adjustment of baseline covariates in a CRT setting is often considered for additional precision gains. Fortunately, many methods are available to estimate adjusted intervention effects in CRT. Examples include well-established methods, such as generalized estimating equations (GEE) and covariate adjusted residuals estimator (CARE), as well as more recent developments, such as targeted maximum likelihood estimation (TMLE) and augmented GEE (Aug-GEE) [19, 41, 108, 114, 115]. While these methods differ in their exact implementation, each aims to improve statistical power in a CRT by controlling for individual or cluster-level covariates when fitting the "outcome regression": the conditional expectation of the outcome, given the randomized intervention and the adjustment variables.

To preserve Type I error control, the adjustment strategy must be specified *a priori*. However, selecting the optimal approach can be especially challenging when there are few clusters and many baseline predictors of the outcome. In small trials, adjusting for too many variables can lead to overfitting and misleading inference. Data-adaptive approaches, which may help overcome this challenge, include backward stepwise regression [116] and adaptive pre-specification [42]. However, common implementations of stepwise regression run a risk of overfitting, and adaptive prespecification is not yet widely used. Previous literature [44, 46] has used simulation studies to compare the power and type I error of different CRT analysis approaches, including: cluster-level methods, GEE, random effects models, 'design-effect' models, and methods ignoring the cluster effect altogether. However, to the best of our knowledge, these comparisons have excluded the more recent approaches of Aug-GEE and TMLE.

This paper aims to clarify the strengths and weaknesses of standard and recently developed methods to analyze CRTs. To motivate this comparison, we consider the Preterm Birth Initiative (PTBi) study, a maternal-infant CRT which took place in 20 health facilities across Kenya and Uganda (NCT03112018). The trial assessed whether a facility-based intervention, designed to improve uptake of evidence-based practices, was effective in reducing 28-day mortality among preterm infants. In Section 2, we use structural causal models to describe the data generating process for a CRT [117]. We also discuss several effects commonly targeted in CRTs. In Section 3, we describes each estimation method, including the statistical properties and assumptions on which it relies. In Section 4, we provide a simulation study constructed to reflect the PTBi study, and we apply these methods to estimate intervention effects for the PTBi study in Section 5. We conclude in Section 6 with a brief discussion.

4.2 Causal Methods

We begin by formalizing notation that will be used throughout. We assume that each cluster, such as a hospital or smaller "health facility" is sampled from some target population of interest and is indexed by $j = 1, ..., J$. Each cluster is comprised of a finite set of individuals (e.g., patients), which are indexed $i = 1, ..., N_j$. The cluster size, denoted N_j, could be

fixed or could vary by cluster. In each cluster, we collect baseline covariates for all the individuals $\mathbf{W}_j = (W_{1j}, \ldots, W_{N_j j})$. We also collect baseline covariates relating to the cluster E_j; these could be aggregates of individual-level characteristics or may have no individual-level counterpart. (We refer the reader to [43] for a discussion of the inclusion of cluster-level summaries of $\mathbf{W}$ in E). In a CRT, the intervention is randomized among clusters. In the case of a pair-matched trial, which may increase the power of the study, the randomization occurs within pairs of clusters [108, 118, 119].

Throughout, we will use A_j as an indicator of whether cluster j was randomized to the intervention ($A_j = 1$), or control ($A_j = 0$). The outcome of interest is $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{N_j j})$, which for simplicity we assume is measured for all individuals in cluster j. Extensions to handling missing data are important, but beyond the scope of this paper [120, 121].

Throughout, we assume that clusters are independent, and we explore varying levels of dependence within a cluster.

Hierarchical Structural Causal Models

We now use Pearl's structural causal model [117] to represent the hierarchical data-generating process of a CRT. These models formally represent the relationships between intervention, outcome, measured characteristics, and unmeasured variables. First, we briefly review the hierarchical model, proposed in [43], in addition to a restricted hierarchical model that more accurately represents the PTBi CRT.

As detailed in [43], the following model makes no restrictions on the dependence of obervations within a cluster.

$$\begin{aligned} E &= f_E(U_E) \\ \mathbf{W} &= f_{\mathbf{W}}(E, U_{\mathbf{W}}) \\ A &= f_A(U_A) \\ \mathbf{Y} &= f_{\mathbf{Y}}(E, \mathbf{W}, A, U_{\mathbf{Y}}) \end{aligned} \tag{4.1}$$

Here, $(f_E, f_{\mathbf{W}}, f_A, f_{\mathbf{Y}})$ denote the set of functions that determine the values of the observed data variables: the cluster-level covariates E, the individual-level covariate matrix $\mathbf{W}$, the intervention assignment A, and the vector of individual-level outcomes $\mathbf{Y}$. These functions are assumed to be common across j, and each accounts for unmeasured factors: $(U_E, U_{\mathbf{W}}, U_A, U_{\mathbf{Y}})$. By design in a CRT, the unmeasured factors determining the intervention assignment U_A are independent of other unmeasured factors. This model assumes clusters are independent, but encodes the many sources of dependence within a cluster. For example, the joint error term $U_{\mathbf{Y}}$ induces correlation among participants' outcomes within a cluster. Furthermore, within cluster j, an individual's outcome Y_{ij} may depend on the covariates of others in the same cluster $\mathbf{W}_j$ but not on the covariates of individuals belonging to other clusters $\mathbf{W}_{j'}$ for $j' \neq j$.

We now consider the data generating process for the PTBi study. We assume each health facility, representing the cluster, is sampled from a target population of interest. For facility

j, we measure facility-level baseline characteristics E_j, including the average monthly delivery volume, facility preparedness assessment score, staff to delivery ratio, and community-type (i.e., urban versus rural). The facility is then randomly assigned to intervention or control A_j. When a mother enters the facility to deliver her baby, covariates for the mother-baby dyad W_{ij} are collected. These include the mother's characteristics, such as age, last menstrual period, and health insurance type, and the baby's characteristics, such as sex, weight, length, and arm circumference. Finally, the baby's vital status is recorded; Y_{ij} is an indicator of infant death within 28 days. Over the course of study follow-up, we observe many such births, but for the primary outcome of interest, we restrict to N_j pre-term births, defined as born before 37 weeks of gestation. This process is repeated for sample size of J facilities.

In the PTBi study, it is reasonable to assume that one mother-baby's covariates are independent from another's, and that one baby's outcome does not impact another's. In other words, the causal dependence is more restricted than in Model 4.1. Therefore, we propose the following model to represent the hierarchical data-generating process for facility j:

$$\begin{aligned} E &= f_E(U_E) \\ A &= f_A(U_A) \\ W_i &= f_W(E, U_{W_i}) \qquad \text{for } i = 1, ..., N_j \\ Y_i &= f_Y(E, W_i, A, U_{Y_i}) \quad \text{for } i = 1, ..., N_j \end{aligned} \tag{4.2}$$

Now we assume the functions are common across ij. We also assume U_{W_i} is independent U_{W_j} and U_{Y_i} is independent of U_{Y_j} for $i \neq j$. We also assume each individual's outcome Y_{ij} is drawn from a common distribution, and this outcome does not depend on the measured covariates of all other individuals in that cluster $\mathbf{W}_j$. This model may be more appropriate when outcomes are not socially or biologically transmitted, as in the real-data example for PTBi.

Counterfactuals and Target Causal Effects

We generate counterfactual outcomes by replacing the structural equation f_A in the causal model with our desired intervention. Let $Y_{ij}(a)$ be the counterfactual outcome for individual i in cluster j if, possibly contrary-to-fact, the individual's cluster received $A = a$. In the PTBi study, $Y_{ij}(1)$ is 28-day vital status for infant i if her or she had been delivered at a facility randomized to the intervention arm $A = 1$, while $Y_{ij}(0)$ is 28-day vital status for infant i if he or she had been delivered at a facility randomized to the control arm $A = 0$.

We can also define counterfactual outcomes at the cluster-level by taking aggregates of the individual-level ones. While many summary measures are possible, we focus on weighted sums of the N_j individuals from cluster j for simplicity:

$$Y_j^c(a) \equiv \sum_{i=1}^{N_j} \alpha_{ij} Y_{ij}(a) \tag{4.3}$$

for some weights α_{ij} for $i = \{1, \ldots, N_j\}$. For example, if we select the weights to be the inverse of the cluster size ($\alpha_{ij} = 1/N_j$) in PTBi, then $Y_j^c(a)$ would be the counterfactual cumulative incidence of death by 28-days if, possibly contrary to fact, j received treatment-level $A = a$. We can then summarize these cluster-level counterfactuals with the empirical mean over the sampled clusters [122–126].

$$\frac{1}{J}\sum_{j=1}^{J} Y_j^c(a)$$

As the number of clusters $J \to \infty$, this converges to the expectation over the population:

$$\Phi^c(a) \equiv \mathbb{E}[Y^c(a)] \tag{4.4}$$

In words, $\Phi^c(a)$ is the expected cluster-level outcome if all clusters in the population had been assigned to treatment-level $A = a$. For the PTBi study, $\Phi^c(a)$ represents the expected incidence of 28-day mortality among preterm infants, if all health facilities had been assigned to intervention arm $A = a$. To define causal effects, we take contrasts of these expected counterfactual outcomes. For the relative effects, we may be interested in the causal risk ratio: $\Phi^c(1)/\Phi^c(0)$. For absolute effects, we may be interested in the causal risk difference (a.k.a., the average treatment effect): $\Phi^c(1) - \Phi^c(0)$. For simplicity we refer to these contrasts as "cluster-level effects".

When defining the cluster-level counterfactual outcome (Eq. 4.3), we can consider various weighting schemes. Selecting the weights as $\alpha_{ij} = 1/N_j$ yields a cluster-level effect giving equal weight to each *cluster*, regardless of its size. However, alternative weighting schemes, yielding different causal effects, may also be of interest. In particular, we can assign equal weight to *individuals*, rather than clusters, by using weights $\alpha_{ij} = J/N_T$, where $N_T \equiv \sum_j N_j$ is the total number of participants (e.g., mother-baby dyads in all health facilities). Now consider the empirical mean of this cluster-level counterfactual over the sampled clusters:

$$\frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\frac{J}{N_T}Y_{ij}(a) = \frac{1}{N_T}\sum_{j=1}^{J}\sum_{i=1}^{N_j}Y_{ij}(a)$$

As sample size grows, this converges to the expectation over the population of clusters, each containing a finite number of individuals:

$$\Phi(a) \equiv \mathbb{E}[Y(a)] \tag{4.5}$$

In words, $\Phi(a)$ is the expected individual-level outcome if all clusters in the population had received treatment-level $A = a$. In the PTBi study, $\Phi(a)$ is the counterfactual risk of mortality for a preterm infant if all health facilities had been assigned to intervention arm $A = a$. As before, we can take the difference or ratio of these expected counterfactual individual-level outcomes to define "individual-level" effects. We note that in general, $\Phi^c(a) = \Phi(a)$ only when each cluster is comprised of the same number of individuals: $N_j = n, \forall j$.

Identifiability

For a randomly sampled cluster, the observed data are the set of measured cluster-level covariates, the matrix of individual-level covariates, the randomized treatment, and the vector of individual-level outcomes:

$$O = (E, \mathbf{W}, A, \mathbf{Y}).$$

As before, we can summarize the outcome vector $\mathbf{Y}$ in a variety of ways. To match the definition of our cluster-level causal parameters, we again consider weighted sums:

$$Y_j^c \equiv \sum_{i=1}^{N_j} \alpha_{ij} Y_{ij} \tag{4.6}$$

We assume the observed data were generated by sampling J times from a data-generating process compatible with either of the causal models described above. This provides a link between the causal model and the statistical model, which is the set of possible distributions of the observed data [19]. Both causal models (Eq. 4.1 and 4.2) encode that the cluster-level treatment is randomized and thus imply a semi-parametric statistical model $\mathcal{M}$. We denote the true underlying distribution that generated the observed data as $\mathbb{P}_0$. Thus, we have J independent, identically distributed (i.i.d.) copies of $O \sim \mathbb{P}_0 \in \mathcal{M}$.

To identify the expected counterfactual cluster-level outcome $\Phi^c(a)$ or the expected counterfactual individual-level outcome $\Phi(a)$ as a function of the observed data distribution, we require the following two assumptions, which are satisfied by design in a CRT. First, there must be no unmeasured confounding, such that $A\mathbf{Y}(a)$. Second, there must be a positive probability of receiving each treatment-level: $\mathbb{P}_0(A = a) > 0$. Given these conditions are satisfied in a CRT, we can express the cluster-level causal parameter $\Phi^c(a) = \mathbb{E}[Y^c(a)]$ as the expected cluster-level outcome under treatment-level of interest $\mathbb{E}_0[Y^c|A = a]$, where subscript 0 is used to denote the observed data distribution $\mathbb{P}_0$; a proof is provided in [43]. Likewise, the individual-level causal parameter $\Phi(a) = \mathbb{E}[Y(a)]$ equals the expected individual-level outcome under treatment-level level of interest: $\mathbb{E}_0[Y \mid A = a]$.

We can gain efficiency in CRTs by adjusting for baseline covariates [40, 42, 108, 113]. Specifically, the cluster-level estimand can be written as the conditional expectation of the cluster-level outcome, given the treatment-level of interest and the baseline covariates, averaged over the covariate distribution: $\mathbb{E}_0[\mathbb{E}_0[Y^c|A = a, E, \mathbf{W}]]$ [43]. In practice, we cannot adjust for the entire matrix of individual-level covariates $\mathbf{W}$ when estimating $\mathbb{E}_0[\mathbb{E}_0[Y^c|A = a, E, \mathbf{W}]]$. Instead, we can improve precision by considering summary measures of this matrix as cluster-level covariates W^c. Then, the cluster-level estimand can be written as

$$\Psi_0^c(a) \equiv \mathbb{E}_0[\mathbb{E}_0[Y^c|A = a, E, W^c]] \tag{4.7}$$

Similarly, the individual-level estimand can be written as the conditional expectation of the individual-level outcome, given the treatment-level of interest and baseline covariates and then averaged over the covariate distribution:

$$\Psi_0(a) = \mathbb{E}_0[\mathbb{E}_0(Y|A = a, E, W)] \tag{4.8}$$

recognizing the slight abuse to notation because the observed data $O \sim \mathbb{P}_0$ are the cluster-level. Here, W denotes the set of covariates for a given individual. We emphasize that covariate adjustment is being used for efficiency gains only and not to control for confounding.

In the running PTBi example, the cluster-level estimand Ψ^c represents the expected *incidence* of 28-day preterm infant mortality in facilities that received intervention-level $A = a$. The individual-level estimand Ψ represents the *risk* of 28-day mortality among preterm infants who were born in facilities that received intervention-level $A = a$.

As before, we can take contrasts of the cluster-level or individual-level parameters to define effects. To give context for the methods comparison, we will focus on relative scale for the remainder of the paper. Specifically, the relative effect is identified at the cluster-level as

$$\frac{\Psi_0^c(1)}{\Psi_0^c(0)} \tag{4.9}$$

and at the individual-level as

$$\frac{\Psi_0(1)}{\Psi_0(0)} \tag{4.10}$$

As detailed below, most analytic methods are only able to estimate one of the above statistical parameters, whereas few have the flexibility to yield both.

4.3 Statistical Methods: Estimation

In this section, we compare and evaluate the methods commonly used to analyze CRT in the context of studies with fewer than 30 clusters. We aim to describe their ability to flexibly adjust for baseline covariates to improve precision and thereby statistical power, while preserving Type I error control. Their ability to handle missingnenss is an area of future work.

We broadly consider two classes of estimation methods: approaches based on cluster-level data and approaches using both individual and cluster-level data. Cluster-level approaches immediately aggregate the data to the cluster; therefore, any covariate adjustment must be at the cluster-level. Examples of cluster-level approaches include the t-test, a simple substitution estimator (a.k.a. parametric G-computation [19]), and targeted maximum likelihood estimation (TMLE). These approaches target cluster-level estimands, such as Equation (4.9).

Individual-level approaches allow for individual-level covariate adjustment, an appealing option, as these pair naturally with individual-level outcomes. The individual-level methods we will consider include: generalized estimating equations (GEE), the covariate adjusted residuals estimator (CARE), and hierarchical TMLE. It is important to note each of these estimators will generally imply a different causal parameter and statistical estimand of interest-either at the cluster-level as in Equation (4.9) or at the individual-level as in Equation (4.10).

We now define the notation used throughout this section. Recall the cluster level-outcome Y^c is defined as in Equation (4.6). We denote the conditional expectation of the cluster-level outcome Y^c, given the cluster-level intervention A, cluster-level covariates E, and cluster-level summaries W^c as

$$\mu^c(a, E, W^c) \equiv \mathbb{E}(Y^c|A = a, E, W^c) \tag{4.11}$$

Likewise, we denote the conditional expectation of the individual-level outcome, given the cluster-level intervention A, cluster-level covariates E, and that individual's covariates W as

$$\mu(a, E, W) \equiv \mathbb{E}(Y|A = a, E, W) \tag{4.12}$$

Throughout, we refer to Equation (4.11) and to Equation (4.12) as the cluster-level and individual-level outcome regressions, respectively. The unadjusted expectation of the cluster-level and individual-level outcomes within intervention arm a are defined as $\mu^c(a) \equiv \mathbb{E}(Y^c|A = a)$ and $\mu(a) \equiv \mathbb{E}(Y|A = a)$, respectively.

We denote the cluster-level propensity score as

$$\pi^c(a|E, W^c) \equiv \mathbb{P}(A = a|E, W^c) \tag{4.13}$$

and the individual-level propensity score as

$$\pi(a|E, W) \equiv \mathbb{P}(A = a|E, W) \tag{4.14}$$

We define unadjusted probabilities $\pi^c(a)$ and $\pi(a)$ analogously.

To match the real data example, we focus on a binary individual-level outcome throughout. However, all analytic approaches are equally applicable with count or continuous outcomes. In the PTBi example, recall Y_{ij} is an indicator that infant i in facility j died by 28-days follow-up and N_j is the number of premature infants in facility j. Then, using weights $\alpha_{ij} = 1/N_j$, the cluster-level outcome for facility j, Y_j^c is the cumulative incidence of mortality by 28 days among preterm infants. In the PBTi example, $\mu^c(a)$ represents the expected incidence of 28-day preterm mortality among all facilities in study arm a, whereas $\mu(a)$ represents the individual-level risk of 28-mortality among preterm infants delivered in a facility in study arm a. Lastly, $\pi^c(1)$ denotes the probability a health facility received the intervention, and $\pi(1)$ represents the probability an infant was delivered in a facility assigned to study arm $A = 1$.

Cluster-Level Approaches

Cluster-level approaches obtain point estimates and inference after the individual-level data have been aggregated to the cluster-level [108, 113]. Most commonly, this aggregation is done by taking the empirical mean within each cluster. However, as previously discussed, we can consider several α_{ij} weighting schemes. We also briefly note that modifications to the following methods can be made to account for pair-matching clusters prior to randomization. However, "breaking the match" and analyzing a pair-matched trial, such as PTBi, as if were completely randomized is also a valid approach and the focus of our discussion.

Unadjusted

Once the data are at the cluster-level, a common approach for estimation and inference is based on contrasts of the treatment-specific averages:

$$\hat{\mu}^c(a) = \frac{1}{J}\sum_{j=1}^{J} \frac{\mathbb{1}(A_j = a)}{\hat{\pi}^c(a)} Y_j^c \tag{4.15}$$

where $\hat{\pi^c}(a)$ denotes the unadjusted estimate of the cluster-level propensity score (i.e. the proportion of clusters in the study receiving treatment-level $A = a$). For simplicity for the remainder of the manuscript, we assume the trial has equal allocation of arms, such that $\hat{\pi^c}(a) = 1/2$. Then, if we let $Y_{a,k}^c$ denote the cluster-level outcome for observation $k = \{1, \ldots, J/2\}$ in treatment-arm $A = a$, the treatment specific mean simplifies to $\hat{\mu}^c(a) = \frac{1}{J/2}\sum_{k=1}^{J/2} Y_{a,k}^c$. In a CRT, $\hat{\mu}^c(a)$ provides an unbiased estimator of $\Phi^c(a) = \mathbb{E}[Y^c(a)]$. In the running PTBi example, $\hat{\mu}^c(a)$ represents the average incidence of 28-day infant mortality among facilities that received intervention level a.

With estimates $\hat{\mu}^c(1)$ and $\hat{\mu}^c(0)$, we can obtain a point estimate of the intervention effect by contrasting them on the scale of interest and obtaining inference with a simple t-test. Statistical power may be improved by considering alternative weighting schemes when summarizing individual-level outcomes to the cluster-level; as previously discussed, however, selections of different weights α_{ij} in Equation (4.3) and Equation (4.6) imply different target parameters.

For relative effects (Equation (4.9)), applying the logarithmic transformation is sometimes recommended when the cluster-level summaries are skewed, which may be more common for rate-type outcomes [108]. However, it is important to note that depending on how this transformation is implemented, the resulting target parameter may be the ratio of the geometric means - as opposed to the ratio of the arithmetic means. (Recall for n observations of some variable X, the geometric mean is $\left(\prod_{i=1}^{n} x_i\right)^{1/n}$ whereas the arithmetic mean is $1/n\sum_{i=1}^{n} x_i$). Specifically, suppose we first take the log of the cluster-level outcomes and

then take the average within each arm:

$$\bar{l}_a \equiv \frac{1}{J/2}\sum_{k=1}^{J/2} log(Y^c_{a,k}) = log\left(\prod_{k=1}^{J/2} Y^c_{a,k}\right)^{\frac{1}{J/2}} \tag{4.16}$$

where $Y^c_{a,k}$ denotes the cluster-level outcome for cluster $k = \{1, \ldots, J/2\}$ in treatment arm $A_j = a$. Applying a t-test to the difference in these treatment-specific means $\bar{l}_1 - \bar{l}_0$ (and then exponentiating) targets the ratio of the geometric means. To avoid changing the target of inference, we can instead apply the Delta Method to obtain point estimates and inference for the standard risk ratio (Equation (4.9)). We refer the reader to [127] for more details.

Cluster-level TMLE with Adaptive Pre-specification

Statistical power may be improved by adjusting for cluster-level covariates E and cluster-level summaries of individual-level covariates W^c. Specifically, once the data have been aggregated to the cluster-level, we can proceed with estimation and inference for $\Psi^c_0(a) = \mathbb{E}_0[\mathbb{E}_0[Y^c|A = a, E, W^c]]$, as if the data were i.i.d. Examples of common algorithms include the simple substitution estimators (a.k.a. parametric G-computation), inverse probability of treatment weighting estimators (IPTW), and targeted maximum likelihood estimators (TMLE) [19]. Due to randomization of the intervention, these algorithms will be consistent, even under misspecification of the outcome regression [128]. Since these methods are well-established and implementation is identical to the non-clustered setting, we focus our discussion on alternative implementations that can harness both individual and cluster-level covariates to increase precision. Before doing so, we briefly review the steps of a cluster-level TMLE and discuss challenges in estimation of the cluster-level outcome regression $\mu^c(A, E, W^c)$ or the cluster-level propensity score $\pi^c(a|E, W^c)$ in trials with limited numbers of randomized units.

TMLE are a general class of double robust, semiparametric efficient plug-in estimators. To implement a cluster-level TMLE, we first obtain an initial estimator of the expected cluster-level outcome $\mu^c(A, E, W^c)$. Next, we update this initial estimator $\hat{\mu}^c(A, E, W^c)$ using information contained in the estimated propensity score $\hat{\pi^c}(a|E, W^c)$. Specifically, we define the "clever covariate" as the inverse propensity score for cluster j:

$$\hat{H}^c(a, E_j, W^c_j) = \frac{\mathbb{1}(A_j = a)}{\hat{\pi^c}(a|E_j, W^c_j)}$$

Then on the logit-scale, we regress the cluster-level outcome Y^c_j on the covariates $\hat{H}^c(1, E_j, W^c_j)$ and $\hat{H}^c(0, E_j, W^c_j)$ with the initial estimator $\hat{\mu}^c(A_j, E_j, W^c_j)$ as the intercept. This provides the following targeted estimator, while simultaneously solves the efficient score equation:

$$\hat{\mu}^{c*}(a, E, W^c) = logit^{-1}[logit(\hat{\mu}^c(a, E, W^c)) + \hat{\epsilon}_1\hat{H}^c(1, E, W^c) + \hat{\epsilon}_0\hat{H}^c(0, E, W^c)] \tag{4.17}$$

where $\hat{\epsilon}_1$ and $\hat{\epsilon}_0$ denote the estimated coefficients for $\hat{H}^c(1, E, W^c)$ and $\hat{H}^c(0, E, W^c)$, respectively. Finally, we obtain a point estimate of the treatment-specific mean $\Psi_0^c(a)$ by averaging the targeted predicted of the cluster-level outcomes across the J study units:

$$\hat{\Psi}^{c*}(a) = \frac{1}{J} \sum_{j=1}^{J} \hat{\mu}^{c*}(a, E_j, W_j^c)$$

To recover the desired target effect, we can contrast of these estimates on the scale of interest, such as

$$\frac{\hat{\Psi}^{c*}(1)}{\hat{\Psi}^{c*}(0)}$$

The variance of asymptotically linear estimators, such as the TMLE, may be estimated using the estimator's influence function. These types of estimators enjoy properties that follow from the central limit theorem and and allow us to recover 95% Wald-type confidence intervals. Alternatively, as described in the next section, we can use the cross-validated influence function to estimate the variance [42].

In a CRT, the propensity score $\pi^c(a, E, W^c) = \pi^c(a)$ is known and does not need to be estimated. If the cluster-level propensity score is not estimated, then this approach corresponds to G-computation [128]. However, further gains in efficiency can be achieved through its estimation [40, 128]. If both the cluster-level outcome regression and cluster-level propensity score are consistently estimated, the TMLE will be an asymptotically efficient estimator. However, consistent estimation of the outcome regression is nearly impossible in trials using an *a priori* specified regression model. To improve precision while preserving Type-I error control, we previously proposed "Adaptive Prespecification," a supervised learning approach using sample-splitting choose the adjustment set that maximizes efficiency [42].

In more detail, we prespecify a set of candidate generalized linear models (GLMs) for $\mu^c(a, E, W^c)$ and $\pi^c(a|E, W^c)$. To reduce the potential for over-fitting in small trials, we recommend limiting these GLMs to having only one or two covariates. We also prespecify a cross-validation scheme; for small trials, we recommend leave-one-cluster-out. To measure performance, we prespecify the squared influence function as our loss function. Then we choose the candidate estimator of $\mu^c(a, E, W^c)$ that minimizes the cross-validated variance estimate using the influence function based on the known propensity score (i.e., $\pi^c(a) = 0.5$). We then select the candidate estimator of propensity score $\pi^c(a|E, W^c)$ that further minimizes the cross-validated variance estimate using the influence function *when combined with the previously selected estimator* $\hat{\mu}^c$. The propensity score estimate should not adjust for a covariate that was already used to estimate $\mu^c(A, E, W^c)$ Together, the selected estimators $\hat{\mu}^c(a, E, W^c)$ and $\hat{\pi}^c(a|E, W^c)$ form the "optimal" TMLE according to the principle of empirical efficiency maximization [40] To account for the data-adaptive procedure used to select this TMLE, we can use a cross-validated influence function for inference. However, with very few candidates, the cross-validated inference may be overly conservative and the non-cross-validated inference may be preferable [42].

Lastly, to incorporate a pair-matched design, we consider the pair as the independent unit for variance estimation and sample splitting; specifically, we recommend leave-one-pair-out cross-validation.

Individual-level Approaches

We now discuss how to leverage individual-level covariates when estimating effects in CRTs. This is can done by aggregating to the cluster-level *after* estimating the expected individual-level outcome, or implementing a fully individual-level approach and accounting for clustering during variance estimation. These types of estimators may have the flexibility to estimate both cluster-level effects (Equation (4.9)) and individual-level effects (Equation (4.10)).

Hierarchical TMLE

In Section 4.3, we discussed a cluster-level TMLE for estimating effects in CRTs based on aggregating the data to the cluster-level and applying the standard TMLE for I.I.D. data. We now present two TMLEs, which account for the hierarchical nature of the data while leveraging the natural pairing of individual-level outcomes with individual-level covariates [43]. Both methods extend to a pair-matched design [42, 129, 130].

Hybrid TMLE:

Recall that the first step of the cluster-level TMLE is to obtain an initial estimator of the conditional expectation of the cluster-level outcome $\mu^c(A, E, W^c)$. Instead of only considering cluster-level approaches, we can expand our candidate estimators by including aggregates of individual approaches [43]. This allows us to harness the pairing of individual-level covariates with individual-level outcomes. Consider, for example, the following specification of the expected individual-level outcome $\mathbb{E}(Y|A, E, W)$:

$$\mu(A, E, W) = logit^{-1}[\beta_0 + \beta_A A + \beta_E E + \beta_W W]$$

We could estimate the coefficients in this regression by pooling individuals across clusters. Afterwards, for each cluster j, we would obtain and average the individual-level predicted outcomes to generate a candidate estimator of the expected cluster-level outcome:

$$\hat{\mu}^c(A_j, E_j, W_j^c) = \sum_{i=1}^{N_j} \alpha_{ij} \times \hat{\mu}(A_j, E_j, W_{ij})$$

Then estimation and inference would proceed as described in Section 4.3 for the cluster-level TMLE. Thus, this approach naturally targets the cluster-level treatment-specific mean $\Psi_0^c(a) = \mathbb{E}_0[\mathbb{E}_0(Y^c|a, E, W^c)]$, and therefore cluster-level effects, as in Equation (4.9). However, as previously discussed, we can set the weights α_{ij} to give equal weight to clusters or to individuals.

More importantly, we can now use Adaptive Prespecification to choose between candidate estimators of μ^c based on cluster-level approaches or aggregates of individual-level approaches. In the latter, the initial estimator of μ^c is based off of pooling individuals across clusters; therefore, we could consider methods that are more data-adaptive than GLMs. Specifically, we could use Super Learner, an ensemble machine learning algorithm [21]. As before, we note that if the cluster-level propensity score is not estimated $\pi^c(a) = 0.5$, then this approach corresponds to G-computation.

Individual-level TMLE:

Instead of only considering candidate individual-level estimators of the expected outcome, we could implement the entire TMLE algorithm at the individual-level and account for clustering when obtaining inference [43]. Specifically, we now consider initial estimators of the individual-level outcome $\mu(A, E, W)$, which are updated using individual-level propensity score $\hat{\pi}(a|E, W)$, instead of the cluster-level propensity score. Both estimators $\hat{\mu}$ and $\hat{\pi}$ would be fit pooling individuals across clusters. As before, we calculate the "clever covariates", but now at the individual-level:

$$\hat{H}(a, E_j, W_{ij}) = \frac{\mathbb{1}(A_j = a)}{\hat{\pi}(a|E_j, W_{ij})}$$

for $a = \{1, 0\}$. Then on the logit-scale, we would regress the individual-level outcome Y_{ij} on the individual-level covariates $\hat{H}(1, E_j, W_{ij})$ and $\hat{H}(0, E_j, W_{ij})$ with the initial individual-level estimator $\mu(A_j, E_j, W_{ij})$ as the intercept. This provides the following updated estimator of the expectation of the individual-level outcome, while simultaneously solves the efficient score equation:

$$\hat{\mu}^*(a, E, W) = logit^{-1}[logit(\hat{\mu}(a, E, W)) + \hat{\epsilon}_1 \hat{H}(1, E, W) + \hat{\epsilon}_0 \hat{H}(0, E, W)] \tag{4.18}$$

where $\hat{\epsilon}_1$ and $\hat{\epsilon}_0$ now denote the estimated coefficients for $\hat{H}(1, E, W)$ and $\hat{H}(0, E, W)$. Then we would obtain a point estimate by averaging these targeted predictions:

$$\hat{\Psi}^*(a) = \frac{1}{N_T} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \hat{\mu}^*(a, E_j, W_{ij})$$

where N_T denotes the total number of participants. Thus, this approach naturally targets the individual-level treatment-specific mean $\Psi_0(a) = \mathbb{E}_0[\mathbb{E}_0(Y|a, E, W)]$, and therefore the individual-level effect, as in Equation (4.10). As detailed in Balzer et al. [43], we can also target the cluster-level effect (Equation (4.9)) by incorporating the weights α_{ij} in each step.

For any choice of weights α_{ij}, statistical inference must respect the cluster as the independent unit. To do so, we aggregate the individual-level influence function to the cluster-level and then take the sample variance of the estimated cluster-level influence function, scaled by J. As before, we can use Adaptive Prespecification to select among candidate estimators of the individual-level expected outcome μ and propensity score π. We may also consider leveraging the larger effective sample size in the Individual-level TMLE using machine-learning methods to μ and π [43].

Covariate-adjusted Residuals Estimator (CARE)

The covariate-adjusted residuals estimator (CARE) was first proposed in [115] and later popularized by [108]. In CARE, we first regress the individual-level outcome on individual-level and cluster-level covariates of interest, but not the cluster-level intervention. Then the predictions from this regression are aggregated to the cluster-level. Finally, a t-test comparing mean residuals (i.e., the discrepancies between observed and predicted outcomes) by arms is performed, since, under the null hypothesis, the average residuals should be the same between arms. This estimator's target estimand is cluster-level effect, such as Equation (4.9) for the relative scale.

CARE is similar to the Hybrid TMLE in that individual-level adjustment occurs before aggregating data to the cluster and proceeding with a cluster-level analysis. However, CARE relies on a fixed specification the individual-level outcome regression that excludes the treatment [115]. For the relative risk (Equation (4.9)) and a binary outcome Y_{ij}, we could, for example, fit the following individual-level logistic regression

$$\mathbb{E}(Y|E,W) = logit^{-1}[\alpha + \beta_E E + \beta_W W] \tag{4.19}$$

where β_E and β_W denote the magnitude by which the log odds of the outcome for the ith individual in the jth cluster is affected (linearly) by either cluster-level covariates E_j or individual-level covariates W_{ij}, respectively. Once this model is fit, the expected number of events in the jth cluster is calculated as $e_j = \sum_i^{N_j}(\hat{\alpha} + \hat{\beta_E} E_j + \hat{\beta_W} W_{ij})$, and compared the observed number of events $d_j = \sum_i^{N_j} Y_{ij}$ through ratio-residuals:

$$R_j = \frac{d_j}{e_j}$$

Hayes and Moulton [108] note that these ratio-residuals are often right-skewed and recommend a logarithmic transformation. Specifically, they recommend applying a t-test to obtain point estimates and inference for the difference in the treatment-specific averages of the log-transformed residuals:

$$\frac{1}{J/2}\sum_{k=1}^{J/2} log(R_{1,k}) - \frac{1}{J/2}\sum_{k=1}^{J/2} log(R_{0,k}) = log\left(\prod_{k=1}^{J/2} R_{1,k}\right)^{\frac{1}{J/2}} - log\left(\prod_{k=1}^{J/2} R_{0,k}\right)^{\frac{1}{J/2}} \tag{4.20}$$

where $R_{a,k}$ denotes the ratio-residual for cluster $k = \{1, \ldots, J/2\}$ in arm a. As detailed in Section 4.3, after exponentiation, we recover estimates and inference for the ratio of the geometric means and thereby a different target parameter than the standard risk ratio, given in Equation (4.9). A straightforward extension to pair-matched design is illustrated in [108].

Generalized estimating equations (GEE)

We now consider a class of estimating equations, sometimes referred to as "population-average models", for estimating effects in CRTs [131]. In GEE, estimation and inference

is conducted at the individual-level and a working correlation matrix is used to account for the dependence of outcomes within clusters. Therefore, the target of inference is now the individual-level effect as in Equation (4.10), instead of a cluster-level effect as in Equation (4.9).

In GEE, the expected individual-level response is modeled a function of the treatment and possibly covariates of interest, but not on any random effects [45, 131]. Specifically, consider the following "marginal model" for the expected individual-level outcome $\mathbb{E}(Y|A)$:

$$\mu(A_j) = g^{-1}(\beta_0 + \beta_A A_j) \tag{4.21}$$

where $g^{-1}(\cdot)$ denotes the inverse-link function. Commonly, the identity link is used for continuous outcomes, the log-link for count outcomes, and the logit-link for binary outcomes. As usually implemented, GEE estimates the intervention effects by examining the point estimates and confidence intervals for the treatment coefficient β_A. At the individual-level, β_A represents causal risk difference for the identity link, e^{β_A} represents the causal risk ratio for the log-link, and e^{β_A} represents the causal odds ratio for the logit-link. In other words, the link function often determines the scale on which the effect is estimated.

As with other CRT approaches, GEE may improve efficiency by adjusting for covariates. Consider, for example, the following "conditional model" for the expected individual-level outcome $\mathbb{E}(Y|A, E, W)$:

$$\mu(A_j, E_j, W_{ij}) = g^{-1}(\beta_0 + \beta_A A_j + \beta_E E_j + \beta_W W_{ij}) \tag{4.22}$$

where again $g^{-1}(\cdot)$ denotes the inverse-link function. Except for linear and log-linear models without interaction terms, the interpretation of β_A is generally not the same as in the marginal model. For the logistic link function, for example, β_A in Equation (4.22) would yield the conditional log-odds ratio instead of the marginal log-odds ratio. However, a recent modification to GEE, presented next, allows for estimation of marginal effects, while adjusting for individual-level or cluster-level covariates.

For either a marginal or conditional specification, the GEE estimator solves the following equation:

$$\sum_{j=1}^{J} \mathbf{D}_j^T \mathbf{V}_j^{-1} (\mathbf{Y}_j - \boldsymbol{\mu}_j) = 0$$

where $\boldsymbol{\mu}_j$ is the vector containing individual-level outcome regressions for cluster j, and $\mathbf{D}_j = \frac{\delta \boldsymbol{\mu}_j}{\delta \beta}$ is the gradient matrix. $\mathbf{V}_j$ is the working correlation matrix used to account for dependence of individuals within a cluster j [131].

GEE yield a consistent estimate of β_A under the marginal model, even if the correlation matrix has been misspecified. However, under misspecification of the covariance matrix, the usual standard errors obtained are not valid, and the sandwich variance estimator must be used [131]. In general, unless the number of clusters is relatively large and the number of participants within cluster is relatively small, the sandwich based standard errors can underestimate the true variance of β, and yield confidence intervals with coverage probability

below desired nominal level. Corrections exist to account for underestimates of low variance when the number of clusters yields fewer than 38 degrees of freedom. [113, 132].

The extension to a pair-matched analysis requires specifying a fixed effect for the pair while maintaining the correlation structure at the cluster-level. However, this is discouraged for studies with fewer than 40 clusters. [113, 132].

Augmented-GEE

As discussed in the previous subsection, the coefficient β_A resulting from GEE does not always correspond to the marginal effect. Recently, a modification to GEE was proposed to ensure β_A can be interpreted as a marginal effect while adjusting for baseline covariates to improve efficiency [41]. This approach is referred to as Augmented-GEE (Aug-GEE) and again targets individual-level effects, as in Equation (4.10).

As commonly implemented, Aug-GEE modifies GEE by including an additional "augmentation" which incorporates the conditional expectation of the individual-level outcome $\mu(A_j, E_j, W_{ij})$. The general form of the Aug-GEE for a binary treatment is given by

$$\sum_{j=1}^{J} \Big[\mathbf{D}_j^T \mathbf{V}_j^{-1} (\mathbf{Y}_j - \boldsymbol{\mu}_j(A_j)) - \sum_{a=0}^{1} \big[\mathbb{1}(A_j = a) - \pi^c(a) \big] \gamma_a \Big] = 0 \tag{4.23}$$

with augmentation term

$$\gamma_a = \mathbf{D}_j^T \mathbf{V}_j^{-1} (\boldsymbol{\mu}_j(a, E_j, W_{ij}) - \boldsymbol{\mu}_j(a))$$

Here, for cluster j, $\boldsymbol{\mu}_j(A_j)$ denotes the vector of marginal regressions (as in Equation (4.21)) and $\boldsymbol{\mu}_j(A_j, E_j, W_{ij})$ denotes the vector of conditional regressions (as in Equation (4.22)). The cluster-level propensity score, defined in Equation (4.13), is treated as known (i.e., $\pi^c(a) = 0.5$). By solving Equation (4.23), we can obtain point estimates and inference for the marginal effects. As with the other estimators considered, if the conditional outcome regression $\mu(A_j, E_j, W_{ij})$ is misspecified, the resulting estimator is asymptotically normal and consistent; however, it is not efficient [41].

The efficiency of the estimator highly depends on the matrix $\mathbf{D}^T \mathbf{V}^{-1}$. Stephens et al. [133] show how to further improve the Aug-GEE by deriving the semiparametric locally efficient estimator, and conclude the derivation of the high dimensional inverse covariance matrix presents a substantial barrier to any additional gains.

4.4 Simulation Study

We simulated a simplified data-generating process to reflect the underlying hierarchical data structure of the PTBi study, which randomized $J = 20$ clusters (health facilities). For each of the 20 clusters, we sampled cluster-level covariate $E_1 \sim Norm(2, 1)$ and $E_2 \sim Norm(0, 1)$ and within-cluster (finite) sample size $N_j \sim Norm(190, 10)$. For each cluster j, we also

simulated the random variable $U_E \sim Unif(-0.2, 1.5)$ which represents an unmeasured source of dependence within each cluster. To reflect the PTBi study design and examine the impacts of "breaking the match" during the analysis, we paired clusters on E_2 using the nonbipartite matching algorithm [134]. Within the pair, one cluster was randomized to the intervention arm and the other to the control arm.

For participant $i = \{1, \ldots, N_j\}$ in each cluster j, we generated two individual-level covariates: $W_1 \sim Norm(4U_E, 0.9)$ and $W_2 \sim Norm(2U_E, 0.35)$. Lastly, we simulated the individual-level outcomes as a function of the intervention A_j, cluster and individual-level covariates, and unmeasured factor $U_{Y_{ij}} \sim Unif(0, 1)$:

$$Y_{ij} \sim \mathbb{1}\left[U_{Y_{ij}} < logit^{-1}(\beta_A A + 0.4W_1 + 0.8W_2 + 0.3E_1 + 0.6AW_2)\right] \tag{4.24}$$

We set the coefficient for the intervention $\beta_A = 0.6$ in order to assess the estimator's power when there was an effect, and $\beta_A = 0$ to assess Type I Error under the null. For both settings, we generated counterfactual outcomes under the intervention and under the control by setting $A = 1$ and $A = 0$, respectively. Then for a population of 20 clusters, we calculated the relative effect at the cluster-level (Equation (4.9)) and at the individual-level (Equation (4.10)).

For our simulation, we assessed the performance of the following estimators: cluster-level estimators (t-test, Cluster TMLE), and individual-level estimators (Hybrid TMLE, Individual-level TMLE, CARE, GEE, and Aug-GEE), each using both an unmatched and pair-matched analysis. The t-test used the mean of the cluster-level empirical incidence in each arm (Equation (4.15)) to estimate Equation (4.9). Each TMLE used adaptive prespecification to choose at most two covariates for adjustment if it improved efficiency relative to the unadjusted estimator. As mentioned in Section 4.3, if a covariate was adjusted for in the outcome regression, it was not considered for estimation of the propensity score. The candidate adjustment set for the Cluster-level TMLE included $\{E_1, E_2, W_1^c, W_2^c\}$ for both the outcome regression and propensity score estimation. The candidate adjustment set for the Hybrid TMLE consisted of $\{E_1, E_2, W_1, W_2, W_1^c, W_2^c\}$ for estimation of the outcome regression, and $\{E_1, E_2, W_1^c, W_2^c\}$ for estimation of the cluster-level propensity score. The Individual-level TMLE could select from $\{E_1, E_2, W_1, W_2\}$ to estimate both the individual-level outcome regression and individual-level propensity score. Inference for TMLE was based on the cross-validated influence function.

CARE and both types of GEE used a fixed specification outcome regression which adjusted for W_2. CARE targeted the ratio of geometric means, and GEE (with the log-link) targeted the estimand in Equation (4.10).

Simulation Results

In Table 4.1, we show the 95% confidence interval coverage, statistical power, and Type I error attained by each estimator (for its corresponding target estimand) across all 2500 iterations. The unadjusted estimator, which serves as a baseline for comparison, maintains nominal Type I error (4%) under the null, but it achieved low statistical power (30%)

with nominal confidence interval coverage (95%). A simple cluster-level adjusted estimator, such as Cluster TMLE, showed increased power (72%) relative to the unadjusted estimator. CARE maintained excellent Type I Error control (5%) and power (76%) in the unpaired case, and lower power using the pair-matched analysis (64%). For this data-generating simulation, both types of GEE methods inflated Type I error (10%-11%) although they achieved high power (85%-86%) (Table 4.1).

Cluster TMLE, Hybrid TMLE, and Individual-level TMLE performed very similarly to each other in terms of confidence interval coverage (96%) and power (72%) for the target estimand Equation (4.9). The data-adaptive approach to covariate adjustment prevented overfitting while maintaining good power. The unmatched Cluster TMLE selected the cluster-level summary W_2^c for adjustment in the cluster-level outcome regression for 34% of the repetitions. The unmatched Hybrid TMLE selected W_2 for adjustment using an individual-level outcome regression for 98% of the repetitions. The unmatched Individual-level TMLE selected W_2 for adjustment in the outcome regression for 98% of the repetitions. All three estimators used the unadjusted propensity score estimator for $93 - 97\%$ of the repetitions in the unmatched analysis. The outcome regression adjustment was similar in the pair-matched analysis. However, in the pair-matched analysis, only about 50% of the propensity score estimations selected the unadjusted estimator across all three TMLE types.

As shown in [42], when the candidate adjustment set is small, the cross-validated influence function may yield overly-conservative variance estimates. Therefore, when considering few covariates for adjustment, we may alternatively use the non-cross validated influence function to obtain a variance estimate.

4.5 Real data application: The PTBi Study in Kenya and Uganda

The PTBi study was a pair-matched CRT to evaluate the impact of a health facility intervention on 28-day mortality among preterm infants in Eastern Uganda and Western Kenya from October 2016 to May 2018 (NCT03112018). The study was motivated by the continued need to address preterm birth as a leading risk factor for perinatal mortality, defined as stillbirth and first-week deaths [53]. Low-tech evidence-based practices (e.g., skin-to-skin contact) are not routinely used in low-middle income countries and have the potential to improve outcomes for preterm infants during the critical intrapartum and immediate newborn periods. Therefore, the study was designed to improve the quality of care surrounding the time of birth for mothers and preterm infants [53].

Twenty public sector health facilities across Western Kenya and Eastern Uganda, consisting of large hospitals and smaller health centers, were pair-matched and randomized either to the intervention or control arm [53]. The facilities ranged in size, staff-patient ratio, and capacity to perform cesarean section (C-section), among other characteristics. Facilities in the control arm received (1) strengthening of routine data collection and data strengthen-

Table 4.1: Estimator performance in data-generating simulation (N=2500)

	Unmatched			Pair-matched		
	covg	power	type I error	covg	power	type I error
t-test	0.952	0.302	0.038	0.953	0.300	0.038
Cluster TMLE	0.962	0.722	0.041	0.962	0.656	0.043
Hybrid TMLE	0.964	0.721	0.040	0.962	0.656	0.043
Ind-level TMLE	0.964	0.723	0.041	0.963	0.649	0.038
CARE	0.954	0.757	0.052	0.952	0.642	0.049
GEE	0.928	0.851	0.096	0.939	0.856	0.107
Aug-GEE	0.922	0.855	0.101	0.941	0.857	0.099

Hierarchical Targeted Maximum Likelihood Estimator (TMLE)
Cluster TMLE: aggregated to cluster and targeted at cluster level (adaptive prespecification)
Hybrid TMLE: aggregated to cluster and targeted at cluster level, with pooled estimates of the outcome regression included as candidate estimators (adaptive prespecification)
Ind-level TMLE: targeting is done at individual level and then targeted outcome is aggregated to cluster (adaptive prespecification)
Covariate-Adjusted Residuals Estimator (CARE)
Generalized Estimated Equations (GEE)
Augmented GEE (Aug-GEE)

ing activities, and (2) introduction of WHO Safe Childbirth Checklist. The facilities in the intervention arm received the components included in the Control Arm, in addition to: (1) PRONTO^{TM} Simulation training [55], and (2) quality improvement (QI) aimed to reinforce and optimize use of evidence-based practices. All study components consist of known interventions and strategies aiming to improve quality of care, teamwork, communication, and data use [53].

The process at each PTBi study site occurred as follows. At each selected study site facility characteristics E_j, which denote the vector of all environmental factors shared by participants in health center j, were measured, for example, $E_j = \{$facility volume, staff ratio, c-section capacity...$\}$. Once each health center j was pair-matched, it was randomized to study arm $A_j = a$, and in the time leading up to the data-collection period, the facility healthcare providers received training in their facility according to the intervention arm they were assigned to. A woman and her infant were exposed to intervention level $A_j = a$ when the woman entered a facility to deliver her infant. The mother's individual-level covariates $W_{mother-ij} = \big\{$age, BMI, SES, smoker, gravidity, parity...$\}$ were collected. Then, for all deliveries listed in maternity registries, infant characteristics $W_{infant-ij} = \big\{$sex, weight, length, APGAR score, head/chest/arm circum$\big\}$ and birth outcome data were captured. Mothers of live preterm infants were approached to enroll in the study for 28-day follow-up. The primary study outcome Y_j^c was the combined incidence of fresh stillbirth and

28-day all-cause mortality among preterm births across health centers, which is estimated by aggregating all individual level outcomes $Y_{ij} = \mathbb{1}\big[\text{preterm infant mortality status at 28-day follow-up}\big]$ in facility j for $i = \{1, \ldots, N_j\}$. We note that infants dying before discharge (stillbirth and predischarge mortality) were also included in this outcome; for these infants, $Y_{ij} = 1$. By the end of the study period, outcomes for a total of N_j mother-infant dyads were observed in each facility. The intraclass correlation coefficient (ICC) in this study was estimated to be 0.009, suggesting a modest level of dependence within clusters [54].

PTBi Analysis & Results

During the data-collection period, an unforeseen political strike led to lack of medical providers at certain facilities, thereby decreasing volume at some facilities while increasing volume at others. The number of preterm births for a given facility (i.e., the cluster size) ranged between 40-366 in the intervention arm, and ranged between 31-447 in the control arm. Differences in cluster size within matched pairs ranged between 9-211. To study the impact of high variability in sample size, we considered two types of weighting schemes: giving equal weight to facilities $\alpha_{ij} = 1/N_j$ and giving equal weight to mother-infant dyads $\alpha_{ij} = J/N_T$.

Table 4.2 shows the unadjusted estimates from cluster-level TMLE, the hybrid TMLE, and the individual-level TMLE - each of which did not adjust for any covariates. All estimates indicated that the intervention reduced the incidence of 28-day mortality among preterm infants. However, effect estimates varied greatly depending on the choice of the weights α_{ij}. This is to be expected given the wide range of cluster sizes and outcomes. For example, in the control arm, the 28-day mortality observed among the two largest facilities was 0.39 among 447 infants and 0.29 among 302 infants. In contrast, in the intervention arm, the 28-day mortality observed among the two largest clinics was 0.23 among 366 infants and 0.06 among 253 infants, respectively.

When weighting facilities equally $\alpha_{ij} = 1/N_j$ (thus targetting the estimand in Equation (4.9)), the risk of 28-day preterm mortality under the intervention (Risk_1) was 0.122, while the risk under the control (Risk_0) was 0.15 (Table 4.2). The corresponding risk ratio (RR) was 0.815 (95% confidence interval (CI): 0.595-1.116). The results were the same fore the 3 TMLEs and were not statistically significant (p=0.174). However, when weighting individuals equally $\alpha_{ij} = J/N_T$ (thus targetting the estimand in Equation (4.10)), results varied by approach and suggested higher mortality risk in both arms. For example, approaches using individual-level data suggested the risk of 28-day preterm mortality under the intervention (Risk_1) was 0.153, while the risk under the control (Risk_0) was 0.233. The corresponding risk ratio of 0.656 (95%CI: 0.531-0.812) and indicated statistical significant reduction in mortality (p=0.002). For comparison, the p-value from the signed rank test for the risk ratio was 0.1933.

In Table 4.3, we study the combined impacts of weighting schemes and covariate adjustment. Due to chance imbalance and despite pair-matching, fewer facilities in the intervention arm had C-section capacity. Therefore, we adjusted for the proportion of mothers receiving a

C-section in cluster-level approaches and an indicator of receiving a C-section in individual-level approaches. For simplicity, we estimated the effect among complete cases: 6 excluded in the intervention arm and 41 excluded in the control arm (an analysis adjusting for missing data will be reported in future work). For all TMLE estimators, we again considered both weighting schemes: $\alpha_{ij} = 1/N_j$ and $\alpha_{ij} = J/N_T$.

All adjusted estimators indicated that the risk of 28-day preterm mortality in the intervention arm (Risk_1) was lower than the risk of 28-day mortality in the control arm (Risk_0; Table 4.3). However, the point estimates and inference varied by weighting scheme and by analytic approach. When weighting clusters equally ($\alpha_{ij} = 1/N_j$), adjusting for C-section did not meaningfully change TMLEs' estimates. Estimates from CARE were most similar to the individual-level TMLE. Effect estimates from GEE and Aug-GEE were the largest. However, given our previous simulation results indicating poor Type I error control for these approaches, we issue caution when interpreting these results.

Adjusting for C-section did substantially change the estimates when weighting individuals equally ($\alpha_{ij} = J/N_T$). After adjustment, the estimated mortality under the control (Risk_0) was smaller and the effect estimate attenuated as compared the unadjusted estimates. Nonetheless, the individual-level TMLE obtained statistically significant results under pair-matching (RR=0.695; p=0.007).

Table 4.2: PTBi results: unadjusted estimators

	Risk_1	Risk_0	RR	CI.low	CI.high	pval
Cluster TMLE	0.122	0.150	0.815	0.595	1.116	0.174
Hybrid TMLE	0.122	0.150	0.815	0.595	1.116	0.174
Ind-level TMLE	0.122	0.150	0.815	0.595	1.116	0.174

(a) $\alpha_{ij} = 1/N_j$

	Risk_1	Risk_0	RR	CI.low	CI.high	pval
Cluster TMLE	0.131	0.267	0.581	0.488	0.693	0.000
Hybrid TMLE	0.153	0.233	0.656	0.531	0.812	0.002
Ind-level TMLE	0.153	0.233	0.656	0.531	0.812	0.002

(b) $\alpha_{ij} = J/N_T$

Paired, unadjusted for PTBi data (N=2938)

Risk_1; Risk of 28-day infant mortality among preterm infants delivered in facility assigned to the intervention arm

Risk_0; Risk of 28-day infant mortality among preterm infants delivered in facility assigned to the control arm

RR; Relative risk

4.6 Discussion

Cluster randomized trials (CRTs) are commonly used to evaluate the effects of interventions, which are randomized to groups of individuals, such as communities, clinics, or schools. In both resource-rich and resource-limited settings, CRTs often have few experimental units (i.e., groups or clusters), which strongly influences statistical power. We compared common CRT methods that aim to improve power by adjusting for cluster- and/or individual-level baseline covariates. We considered each approach theoretically and then using simulations reflecting a small-sample CRT data-generating process, we explored how different analytic strategies impacted performance in terms of statistical power, confidence interval coverage, and type I error. We also applied each method to estimate the intervention effect in the PTBi study. While the estimators performed similarly in simulation, estimates of the PBTi intervention effect varied greatly by analytic approach. This was likely caused by the PTBi study's highly variable cluster sizes, and by a potential interaction between cluster size and treatment. Our findings highlight the importance of distinguishing between individual- and cluster-level estimands, and demonstrate in a real-data example how the weighting scheme can strongly influence effect estimates.

In simulations, unadjusted approaches, as expected, had the lowest statistical power. Methods based on generalized estimating equations (GEE and Aug-GEE) failed to preserve Type I error control; this is consistent with the literature warning against their application (without corrections) with fewer than 30 clusters [108, 113]. The covariate adjusted residuals estimator (CARE) and targeted maximum likelihood estimator (TMLE) both yielded high statistical power, while maintaining nominal confidence interval coverage.

In simulations, there was little difference in performance between cluster-level, hybrid, and individual-level TMLEs; they attained similar confidence interval coverage (96%) and power (72%). We hypothesize this is due to the data-generating mechanism where a single individual-level covariate was strongly predictive of the outcome, but also strongly influenced by a cluster-level covariate. In nearly all iterations (98%), this covariate was selected for adjustment at the individual-level in the hybrid TMLE. All three TMLEs also tended to select an unadjusted estimate of the propensity score. We expect to see larger differences when the cluster size is more variable; in this simulation, the sample sizes per cluster ranged between 150-230. Additional simulations (forthcoming) will incorporate an interaction between the intervention and cluster size.

In the real data application, results greatly varied depending on estimation approach and weighting scheme. Larger facilities in the control arm of PTBi had worse outcomes; therefore, unadjusted estimates of the intervention effect were substantially stronger when targeting individual-level effect (i.e., weighting infants equally) than when targeting a cluster-level effect (i.e., weighting facilities equally). Larger facilities were also more likely to have C-section capacity and thereby more likely to receive higher risk cases. Adjusting for C-section attenuated estimates of the intervention effects in nearly all approaches. Altogether, this real-data example demonstrates of the importance of differentiating between individual-level and cluster-level target estimands.

There are several areas of future work. First, additional study is needed of the performance of all methods when cluster size is "informative" - there is an interaction between cluster size and the intervention [135]. Secondly, in our analysis of the PTBi data, estimation only among complete cases would bias our results if the data are not missing at random; we plan to address missing data in a future analysis. Third, the PTBi analysis only adjusted for C-section, and additional covariates should be considered. Fourth, during the delivery, the infant may have also received additional interventions which could influence their outcome. Unfortunately, this information was not collected in a standardized manner across facilities. Future work could seek a proxy for measuring and studying the impact of the individual-level interventions that occurred in the PTBi study. Lastly, in this analysis, we assumed a one-to-one relationship between mother and infant for simplicity; however, in further sensitivity analyses, we will formally consider accounting for all twins and triplets. Standard error estimates may need to account for this additional layer of clustering. For example, in TMLE, we plan to average the influence function estimate among twins or triplets within a health center.

Overall, this manuscript demonstrates many of the common challenges when analyzing CRTs, despite the wide availability of statistical methods for hierarchical data. CRTs are a popular approach for researchers seeking to scale up interventions from the individual-level to the cluster-level. However, in both resource-rich and resource-limited settings, high costs and barriers to implementation often limit the number of clusters which can be enrolled and randomized. Therefore, it is important to understand the potential advantages and pitfalls of common analytic approaches in small-scale CRTs. Many of the challenges faced by the PTBi study are common in real-world implementations of CRT and highlight the importance of carefully selecting the target estimand and corresponding analytic approach to maximize statistical precision, while maintaining nominal Type I error. We recommend considering both interpretation and impacts on statistical power when selecting the weighting scheme. We also recommend using TMLE given its flexibility to data-adaptively adjust for both cluster-level and individual-level covariates, while preserving Type I error control.

Table 4.3: PTBi results: adjusted estimators

	Risk1	Risk0	RR	CI.low	CI.high	pval
Cluster TMLE	0.128	0.145	0.885	0.585	1.338	0.542
Hybrid TMLE	0.128	0.145	0.885	0.585	1.338	0.542
Ind-level TMLE	0.124	0.149	0.836	0.495	1.411	0.481
GEE	0.121	0.153	0.792	0.425	1.475	0.430
Aug-GEE	0.155	0.227	0.682	0.543	0.856	0
CARE	NA	NA	0.841	0.476	1.484	0.541

(a) Unmatched $\alpha_{ij} = 1/N_j$

	Risk1	Risk0	RR	CI.low	CI.high	pval
Cluster TMLE	0.128	0.145	0.885	0.587	1.335	0.518
Hybrid TMLE	0.128	0.145	0.885	0.587	1.335	0.518
Ind-level	0.124	0.149	0.836	0.600	1.164	0.252
GEE	0.077	0.106	0.724	0.616	0.852	0.000
Aug-GEE	0.080	0.117	0.685	0.571	0.821	0
CARE	NA	NA	0.899	0.589	1.374	0.585

(b) Pair-Matched $\alpha_{ij} = 1/N_j$

	Risk1	Risk0	RR	CI.low	CI.high	pval
Cluster TMLE	0.131	0.167	0.835	0.568	1.227	0.338
Hybrid TMLE	0.131	0.167	0.835	0.568	1.227	0.338
Ind-level	0.148	0.213	0.695	0.435	1.109	0.120

(c) Unmatched $\alpha_{ij} = J/N_T$

	Risk1	Risk0	RR	CI.low	CI.high	pval
Cluster TMLE	0.131	0.167	0.835	0.592	1.178	0.266
Hybrid TMLE	0.131	0.167	0.835	0.592	1.178	0.266
Ind-level TMLE	0.148	0.213	0.695	0.549	0.880	0.007

(d) Pair-Matched $\alpha_{ij} = J/N_T$

Adjusting for individual indicator of C-section performed for PTBi data (among complete cases N=2891)
TMLE variance based on non-cross-validated influence curve
Generalized Estimating Equations (GEE)
Augmented GEE (Aug-GEE)
Covariate-adjusted residuals estimator (CARE)

Chapter 5

Conclusions and future directions

The three studies in this dissertation illustrated the potential utility of data-adaptive methods for informing implementation and uptake of EBP in LMIC. While many approaches in public health research exist to improve uptake of EBP, our emphasis was on two particular approaches: monitoring of individual-level risk to develop targeted interventions, and implementation of CRT. Our analyses used real-data examples based on communities in rural and peri-urban East Africa, and our long-term goal is that these types of research approaches will influence policy-making to reduce the disproportionate burden of disease in LMIC.

The first case study demonstrated an approach for identifying risk of viremia among the UARTO study participants in rural Uganda, and developed a hypothetical differentiated care strategy based on machine learning methods applied to EAM monitoring. When compared to current non-differentiated viral load testing approaches, the hypothetical approach reduced testing frequency while maintaining high sensitivity. The second case study monitored risk of preterm birth in the context of the Preterm Birth Initiative and a peri-urban cohort study in Rwanda, demonstrating excellent prediction of ultrasound-based GA among women who received first-trimester ultrasound. The GA prediction algorithm has implications for a future cost-effective tool for monitoring preterm birth risk in resource-limited settings. Lastly, we demonstrated advantages and pitfalls of several statistical methods used for analyzing CRT with few clusters. We showed that while there is room for precision gains by carefully adjusting for baseline covariates, it is essential to carefully define the statistical parameter of interest and the weighting scheme. We demonstrated how certain methods may help leverage a limited number of clusters to recover precise effect estimates, and illustrated these ideas using a real-data example from the PTBi study in Kenya and Uganda.

Implementation of risk-monitoring algorithms for uptake in routine clinical practice would require integration of the algorithm into user friendly software. Decreasing costs of devices such as tablets and cell phones may facilitate implementation and uptake in routine clinical care [90, 92]. However, the factors limiting uptake of machine learning and data-driven solutions in LMIC require careful consideration. Such methods require robust data systems and accurate data reporting. If models are trained on limited data that is highly prone to error, this can bias algorithms in a damaging way [6]. Additionally, ethical standards to

protect the interests of communities in LMICs need to be in place [35, 36]. Thus, while there is much to be gained by using machine learning in these settings, they must be carefully tailored, and many structures must first be in place such that users and policy makers can actually act on information gained from algorithms in ways that benefit the community.

Future directions of this dissertation should expand the analysis of CRT methods to better understand the role of informative sample size, and should assess comparative performance of the analysis methods under a formal missing data analysis. Both the EAM study and the GA prediction study would greatly benefit from transportability analyses to better validate the potential of such algorithms in different communities within East Africa. Other reliable adherence monitoring methods should be considered and analyzed using a similar machine learning approach, to validate the results obtained using EAM data. This is particularly important because EAM may not reflect routine adherence of persons receiving ART in LMIC [48].

In conclusion, our risk-monitoring analyses showed potential for bridging gaps in EBP uptake through targeted interventions. In future research, it is essential to investigate how to best integrate these approaches into routine clinical practice. Our comparative methods analysis in the context of an EBP-driven CRT demonstrated the importance of CRT for informing future research and policy; these studies require careful analysis, especially under resource constraints. Through the two approaches of risk-monitoring and CRT using data-adaptive methods, public health researchers can be better equipped to meet the needs of stakeholders in resource-constrained settings and optimize delivery of care.

Bibliography

[1] Ulysses Panisset et al. "Implementation research evidence uptake and use for policy-making". In: *Health Research Policy and Systems* 10.20 (2016), pp. 229–237.

[2] United Nations. *The Millennium Development Goals Report 2010*. United Nations, New York, 2011.

[3] Ib Christian Bygbjerg. "Double Burden of Noncommunicable and Infectious Diseases in Developing Countries". In: *Science* 337.6101 (2012), pp. 1499–1501.

[4] Mawuli Komla Kushitor and Sandra Boatemaa. "The double burden of disease and the challenge of health access: Evidence from Access, Bottlenecks, Cost and Equity facility survey in Ghana". In: *PLOS One* 3.3 (2018), e0194677.

[5] Jeffrey A. Claridge M.D. and Timothy C. Fabian M.D. "History and Development of Evidence-based Medicine". In: *World Journal of Surgery* 29 (2005), pp. 547–553.

[6] Eleazar Ndabarora, Jennfier A. Chipps, and Leana Uys. "Systematic review of health data quality management and best practices at community and district levels in LMIC." In: *Information Development* 30.2 (2012), pp. 103–120.

[7] Lisa M. Puchalski Ritchie et al. "Low- and middle-income countries face many common barriers to implementation of maternal health evidence products". In: *Journal of Clinical Epidemiology* 76 (2016), pp. 229–237.

[8] Paul Chinnock, Nandi Siegfried, and Mike Clarke. "Is Evidence-Based Medicine Relevant to the Developing World?" In: *PLOS Medicine* 2.5 (2005), e107.

[9] Alan Pearson and Zoe Jordan. "Evidence-based healthcare in developing countries". In: *International Journal of Evidence-Based Healthcare* 8.2 (2010), pp. 97–100.

[10] Edeghonghon Olayemi, Eugenia V. Asare, and Amma A. Benneh-Akwasi Kuma. "Guidelines in lower-middle income countries". In: *British Journal of Haematology* 177.6 (2017), pp. 846–854.

[11] Elizabeth Leonard, Imke de Kock, and Wouter Bam. "Barriers and Facilitators to implementing evidence-based health innovations in low- and middle-income countries: a systematic literature review". In: *Evaluation and Program Planning* In Press (2020).

[12] Till Bärnighausen et al. "Interventions to increase antiretroviral adherence in sub-Saharan Africa: a systematic review of evaluation studies". In: *Lancet Infectious Disease* 11.12 (2014), pp. 5942–951.

[13] Catherine Barker, Arin Dutta, and Kate Klein. "Can differentiated care models solve the crisis in HIV treatment financing? Analysis of prospects for 38 countries in sub-Saharan Africa". In: *Journal of the International AIDS Society* (2017).

[14] Jessica E. Haberer et al. "Improving antiretroviral therapy adherence in resource-limited settings at scale: a discussion of interventions and recommendations". In: *Journal of the International AIDS Society* (2017).

[15] Zulfiqar A et al. Bhutta. "Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost?" In: *The Lancet* 384.9940 (2014), pp. 347–370.

[16] Fernando C. Barros et al. "Global report on preterm birth and stillbirth (3 of 7): evidence for effectiveness of interventions". In: *BMC Pregnancy and Childbirth* 10.Suppl 1 (2010), p. 53.

[17] Kenneth R. Foster, Robert Koprowski, and Joseph D. Skufca. "Machine learning, medical diagnosis, and biomedical engineering research - commentary". In: *Biomedical Engineering Online* 13.94 (2014), pp. 347–370.

[18] Junfei Qiu et al. "A survey of machine learning for big data processing". In: *EURASIP Journal on Advances in Signal Processing* 67 (2016).

[19] Mark J. van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* 1st edition. Springer Science+Business Media, 2011.

[20] James O. Berger. *Statistical decision theory and Bayesian Analysis.* 2nd. New York: Springer-Verlag, 1985.

[21] Mark van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007).

[22] David H.Wolpert. "Stacked generalization". In: *Neural Networks* 5.2 (1992), pp. 241–259.

[23] Mark van der laan and Sandrine Dudoit. "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment". In: *Statistical Methodology* 2.2 (2005), pp. 131–154.

[24] Aad van der Vaart and Sandrine Dudoit Mark van der laan. "Oracle inequalities for multi-fold cross-validation". In: *Statistics and Decisions* 24.3 (2006), pp. 351–371.

[25] Wenjing Zheng et al. "Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies". In: *Statistics in Medicine* (2017).

[26] Frank J. Schwebela and Mary E. Larimera. "Using text message reminders in health care services: A narrative literature review". In: *Internet Interventions* 13 (2018), pp. 82–104.

[27] Jessica E. Haberer, Nicholas Musinguzi, and Alexander C. Tsai. "Real-time electronic adherence monitoring plus follow-up improves adherence compared with standard electronic adherence monitoring". In: *AIDS* (2017).

[28] *Seventy-first world health assembly. Digital Health.* World Health Organization. May 2018. URL: `https://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_R7-en.pdf`.

[29] Khansoudaphone Phakhounthong et al. "Machine learning-based in-line holographic sensing of unstained malaria-infected red blood cells". In: *Journal of Biophotonics* 11.9 (2018), e201800101.

[30] Khansoudaphone Phakhounthong et al. "Predicting the Severity of Dengue Fever in Children on Admission Based on Clinical Features and Laboratory Indicators: Application of Classification Tree Analysis". In: *BMC Pediatrics* 18.109 (2018).

[31] Dong Jiang et al. "Mapping the transmission risk of Zika virus using machine learning models". In: *Acta Tropica* 185 (2018), pp. 391–399.

[32] Subhash Chandir et al. "Using Predictive Analytics to Identify Children at High Risk of Defaulting From a Routine Immunization Program: Feasibility Study". In: *JMIR Public Health Surveillance* 4.3 (2018), e63.

[33] Tao Liu et al. "Optimal Allocation of Gold Standard Testing Under Constrained Availability: Application to Assessment of HIV Treatment Failure". In: *Journal of the American Statistical Association* 108.504 (2013), pp. 1173–1188.

[34] Katelyn J. Rittenhouse et al. "Improving preterm newborn identification in low-resource settings with machine learning". In: *PLOS One* 14.2 (2019), e0198919.

[35] Amy K. Paul and Merrick Schaefer. "Safeguards for the use of artificial intelligence and machine learning in global health". In: *Bulletin of the World Health Organization* 98 (2020).

[36] Nina Schwalbe and Brian Wahl. "Artificial intelligence and the future of global health". In: *Lancet* 395 (2020).

[37] Michael E. Sobel. "What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1398–1407.

[38] RL Goldenberg et al. "Routine antenatal ultrasound in low- and middle-income countries: first look - a cluster randomised trial". In: *International Journal of Obstetrics Gynaecology* 125.12 (2018), pp. 1591–1599.

[39] Elizabeth Miller, Kelley A. Jones, and Lisa Ripper. "An Athletic Coach–Delivered Middle School Gender Violence Prevention Program: A Cluster Randomized Clinical Trial". In: *JAMA pediatrics* 174.3 (2020), pp. 241–249.

[40] Daniel B Rubin and Mark J van der Laan. "Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis". In: *The International Journal of Biostatistics* 4.1 (2008).

[41] Alisa J. Stephens, Eric J. Tchetgen Tchetgen, and Victor de Gruttola. "Augmented GEE for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster and individual-level covariates". In: *Statistics in Medicine* 31.10 (2012), pp. 915–930.

[42] Laura Balzer, Mark J. van der Laan, and Maya Petersen. "Adaptive pre-specification in randomized trials with and without pair-matching". In: *Statistics in Medicine* (2019).

[43] Laura Balzer et al. "A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure". In: *Statistical Methods in Medical Research* (2018).

[44] Scarlett L. Bellamy et al. "Improved Designs for Cluster Randomized Trials". In: *Statistical Methods in Medical Research* 9 (2000), pp. 135–159.

[45] Alan E. Hubbard et al. "To GEE or Not to GEE: Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health". In: *Epidemiology* 21.4 (2010), pp. 467–474.

[46] Sally Galbraith, James A. Daniel, and Bryce Vissel. "A Study of Clustered Data and Approaches to Its Analysis". In: *Journal of Neuroscience* 30.32 (2010), pp. 10601–10608.

[47] Jessica E. Haberer et al. "Real-Time Adherence Monitoring for HIV Antiretroviral Therapy". In: *AIDS Behav* (2010).

[48] Jeffrey I. Campbell et al. "Ugandan Study Participants Experience Electronic Monitoring of Antiretroviral Therapy Adherence as Welcomed Pressure to Adhere." In: *AIDS Behav* (2018).

[49] Etienne Nsereko et al. "Maternal genitourinary infections and poor nutritional status increase risk of preterm birth in Gasabo District, Rwanda: a prospective, longitudinal, cohort study". In: *BMC Pregnancy and Childbirth* 20.345 (2020).

[50] Sabine Musange et al. "Group antenatal care versus standard antenatal care and effect on mean gestational age at birth in Rwanda: protocol for a cluster randomized control trial." In: *Gates Open Research* 3.1548 (2019).

[51] Jeannette R Ickovics et al. "Cluster Randomized Controlled Trial of Group Prenatal Care: Perinatal Outcomes Among Adolescents in New York City Health Centers". In: *American Journal of Public Health* 106.2 (2016), pp. 359–365.

[52] Felix Sayinzoga et al. "Impact of Group Antenatal Care on Gestational Age at Birth: Results from a cluster randomized control trial in Rwanda". In: (Under review).

[53] Phelgona Otieno et al. "Strengthening intrapartum and immediate newborn care to reduce morbidity and mortality of preterm infants born in health facilities in Migor County, Kenya and Busoga Region, Uganda: a study protocol for a randomized control trial". In: *Annual Review of Public Health* 19.313 (2018).

[54] Dilys Walker et al. "Impact of an intrapartum and perinatal quality improvement package on fresh stillbirth and neonatal mortality among preterm births in Kenya and Uganda; A cluster randomised hospital-based trial". In: *Lancet Global Health* (2020), In press.

[55] Dilys Walker et al. "Impact Evaluation of PRONTO Mexico: A Simulation-Based Program in Obstetric and Neonatal Emergencies and Team Training". In: *Simulation in Healthcare* 11.1 (2015), pp. 1–9.

[56] UNAIDS. "UNAIDS Data 2017". In: *Joint United Nations Programme on HIV/AIDS (UNAIDS)* (2017).

[57] WHO. "Consolidated Guidelines on the use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Pubic Health Approach". In: *World Health Organization* Second Edition 2016 (2016).

[58] Greer Waldrop et al. "Stable patients and patients with advanced disease: consensus definitions to support sustained scale up of antiretroviral therapy". In: *Tropical Medicine and International Health)* 67.2 (2016), pp. 301–320.

[59] Roos E Barth et al. "Rapid accumulation of nonnucleoside reverse transcriptase inhibitor-associated resistance: evidence of transmitted resistance in rural South Africa". In: *AIDS* (2008).

[60] Becky L. Genberg et al. "Patterns of antiretroviral therapy adherence and impact on HIV RNA among patients in North America". In: *AIDS* (2012).

[61] Jessica H. Oyugi et al. "Treatment interruptions predict resistance in HIV-positive individuals purchasing fixed-dose combination antiretroviral therapy in Kampala, Uganda". In: *AIDS* (2007).

[62] Maya Petersen et al. "Long term consequences of the delay between virologic failure of highly active antiretroviral therapy and regimen modification". In: *AIDS* (2008).

[63] Andrew N. Phillips et al. "Cost Effectiveness of Potential ART Adherence Monitoring Interventions in Sub-Saharan Africa". In: *PLOS One* (2016).

[64] Maya Petersen et al. "Super Learner Analysis of Electronic Adherence Data Improves Viral Prediction and May Provide Strategies for Selective HIV RNA Monitoring". In: *JAIDS* (2015).

[65] David W. Haas and Philip E. Tarr. "Perspectives on pharmacogenomics of antiretroviral medications and HIV-associated comorbidities". In: *Curr Opin HIV AIDS* (2015).

[66] Shalom Spira et al. "Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance". In: *Journal of Antimicrobial Chemotherapy* (2003).

[67] Jessica E. Haberer et al. "ART adherence and viral suppression are high among most non-pregnant individuals with early-stage, asymptomatic HIV infection: an observational study from Uganda and South Africa". In: *JIAS; In press* (2018).

[68] Mark J. Siedner. "HIV Infection, Co-Infections, Immune Activations, and Chronic Comorbidities in Rural Uganda". In: *Innovation in Aging* (2017).

[69] Susan A. Adakun et al. "Higher baseline CD4 cell count predicts treatment interruptions and persistent viremia in patients initiating ARVs in rural Uganda". In: *JAIDS* (2013).

[70] Diane V. Havlir et al. "Predictors of Residual Viremia in HIV-Infected Patients Successfully Treated with Efavirenz and Lamivudine plus either Tenofovir or Stavudine". In: *Journal of Infectious Diseases* (2005).

[71] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29 (2001).

[72] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[73] George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Welsey, 1973.

[74] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[75] Erin LeDell, Maya Petersen, and Mark van der Laan. "Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates". In: *Electronic Journal of Statistics* (2015).

[76] *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. Computer Program. 2018. URL: `https://www.R-project.org/`.

[77] Eric C. Polley et al. *SuperLearner: Super Learner Prediction*. Computer Program. 2018. URL: `https://cran.r-project.org/web/packages/SuperLearner/`.

[78] T. Chen, T. He, and M. Benesty. *xgboost: Extreme Gradient Boosting*. Computer Program. 2018. URL: `https://CRAN.R-project.org/package=xgboost`.

[79] A. Kapelner and J. Bleich. "bartMachine: Machine Learning with Bayesian Additive Regression Trees". In: *Journal of Statistical Software* (2016).

[80] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22.

[81] A. Gelman and Y.-S. Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. Computer Program. 2018. URL: `https://CRAN.R-project.org/package=arm`.

[82] Tobias Sing et al. "ROCR: visualizing classifier performance in R. Bioinformatics". In: *Bioinformatics* 21.20 (2005), pp. 3940–3941.

[83] S. Kundu, Y.S. Aulchenko, and A.C.J.W. Janssens. *PredictABEL: Assessment of Risk Prediction Models*. Computer Program. 2020. URL: `https://cran.r-project.org/web/packages/PredictABEL/PredictABEL.pdf`.

[84] Michael Abouyannis, Joris Menten, and Agnes Kiragga. "Development and validation of systems for rational use of viral load testing in adults receiving first-line antiretroviral treatment in sub-Saharan Africa". In: *AIDS* (2011).

[85] Cecilia Ferreyra et al. "Evaluation of clinical and immunological markers for predicting virological failure in a HIV/AIDS treatment cohort in Busia, Kenya". In: *PLOS ONE* 7.11 (2012).

[86] David Meya et al. "Development and evaluation of a clinical algorithm to monitor patients on antiretrovirals in resource-limited settings using adherence, clinical and CD4 cell count criteria". In: (2009).

[87] W.D.F. Venter et al. "Dolutegravir plus Two Different Prodrugs of Tenofovir to Treat HIV". In: *The New England Journal of Medicine* (2019).

[88] Andrew N. Phillips, Deenan Pillay, and Valentina Cambiano. "Effect on transmission of HIV-1 resistance of timing of implementation of viral load monitoring to determine switches from first to second-line antiretroviral regimens in resource-limited settings". In: *AIDS* (2011).

[89] Norma C. Ware et al. "The Meanings in the messages: how SMS reminders and real-time adherence monitoring improve antiretroviral therapy adherence in rural Uganda". In: *AIDS* (2016).

[90] Ramnath Subbaraman. "Digital adherence technologies for the management of tuberculosis therapy: mapping the landscape and research priorities". In: (2018).

[91] Catherine A. Koss et al. "Early Adopters of Human Immunodeficiency Virus Preexposure Prophylaxis in a Population-based Combination Prevention Study in Rural Kenya and Uganda". In: *Clinical Infectious Diseases* ().

[92] Laura Silver and Courtney Johnson. *Internet Connectivity Seen as Having Positive Impact on Life in Sub-Saharan Africa; 1. Majorities in sub-Saharan Africa own mobile phones, but smartphone adoption is modest*. Report. Pew Research Center, 2018.

[93] S. B. McKearnan et al. "Performance of the Net Reclassification for Nonnested Models and a Novel Percentile-Based Alternative". In: *American Journal of Epidemiology* (2018).

[94] Li Liu et al. "Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis". In: *The Lancet* 385.9966 (2015), pp. 430–440.

[95] AMANHI (Alliance for Maternal and Newborn Health Improvement. "Development and validation of a simplified algorithm for neonatal gestational age assessment–protocol for the Alliance for Maternal Newborn Health Improvement (AMANHI) prospective cohort study". In: *Journal of Global Health* 7.2 (2017).

[96] Melissa A. Woythaler, Marie McCormick, and Vincent C. Smith. "Late Preterm Infants Have Worse 24-Month Neurodevelopmental Outcomes Than Term Infants". In: *Pediatrics* 127.3 (2011), e622–e629.

[97] Michael K. Mwaniki et al. "Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: a systematic review". In: *Lancet* 379.9814 (2012), pp. 445–452.

[98] Laura Jeliffe-Pawlowski. "Gestation dating by metabolic profile at birth: a California cohort study". In: *American Journal of Obstetrics and Gynecology* (2016).

[99] P. Sladekevicius et al. "Ultrasound dating at 12–14 weeks of gestation. A prospective cross-validation of established dating formulae in in-vitro fertilized pregnancies". In: *Ultrasound in Obstetric and Gynecology* 26 (2005), pp. 504–511.

[100] David A. Savitz et al. "Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination". In: *American Journal of Obstetrics and Gynecology* 187.6 (2002), pp. 1660–1666.

[101] Imtiaz Jehan et al. "Dating gestational age by last menstrual period, symphysis-fundal height, and ultrasound in urban Pakistan". In: *Ultrasound Obstetric Gynecology* 110 (2010), pp. 231–234.

[102] Robin B. Kalish et al. "First- and second-trimester ultrasound assessment of gestational age". In: *American Journal of Obstetrics and Gynecology* 191.3 (2004), pp. 975–878.

[103] Kenneth Finlayson and Soo Downe. "Why Do Women Not Use Antenatal Services in Low- and Middle-Income Countries? A Meta-Synthesis of Qualitative Studies". In: *PLOS Medicine* 10.1 (2013), e1001373.

[104] Anne Lee et al. "Diagnostic Accuracy of Neonatal Assessment for Gestational Age Determination: A Systematic Review". In: *Pediatrics* 149.6 (2017), e20171423.

[105] Jose Villar et al. "International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project". In: *Lancet* 384.9946 (2014), pp. 857–868.

[106] Jerome H. Friedman. "Multivariate adaptive regression splines (with discussion)". In: *The Annals of Statistics* 19.1 (1991), pp. 1–67.

[107] Leo Breiman et al. *Classification and Regression Trees*. Wadsworth, 1984.

[108] Richard J. Hayes and Lawrence Hale Moulton. *Cluster randomised trials*. 2nd edition. Chapman & Hall/CRC, 2017.

[109] Elizabeth L. Turner et al. "Review of Recent Methodological Developments in Group-Randomized Trials: Part 1—Design". In: *Surveillance* 107.6 (2017).

[110] Elizabeth L. Turner et al. "Review of Recent Methodological Developments in Group-Randomized Trials: Part 2—Analysis". In: *Surveillance* 107.7 (2017).

[111] Catherine M. Crespi. "Improved Designs for Cluster Randomized Trials". In: *Annual Review of Public Health* 37.1 (2016), pp. 1–16.

[112] Christina Pagel et al. "Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications". In: *Trials* 12.151 (2011).

[113] David M. Murray et al. "Design and analysis of group-randomized trials in cancer- A review of current practices". In: *Preventive Medicine* 111 (2018), pp. 241–247.

[114] Kung-Yee Liang and Scott L. Zeger. "Longitudinal Data Analysis Using Generalized Linear Model". In: *Biometrika* 73.1 (1986), pp. 13–22.

[115] Mitchell Gail et al. "On design considerations and randomization-based inference for community intervention trials". In: *Statistics in Medicine* 15 (1996), pp. 1069–1092.

[116] M.A. Efromyson. *Multiple regression analysis*. Ed. by Anthony Ralston and Herbert Wilf. Mathematical Models for Digital Computers. Wiley, 1960, pp. 191–203.

[117] Judea Pearl. *Causality: models, reasoning and inference*. 2nd edition. Cambridge University Press, 2009.

[118] Kosuke Imai, Gary King, and Clayton Nall. "The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation." In: *Statistical Science* 24.1 (2009), pp. 29–53.

[119] Laura B. Balzer et al. "Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation". In: *Statistics in Medicine* 34.6 (2015), pp. 999–1011.

[120] Karla Diaz-Ordaz et al. "Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines". In: *Clinical Trials* 11.5 (2014), pp. 590–600.

[121] Mallorie H. Fiero et al. "Statistical analysis and handling of missing data in cluster randomized trials: a systematic review". In: *Trials* 17.72 (2016).

[122] Jerzy Neyman. "Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990)". In: *Statistical Science* 5 (1923), pp. 463–472.

[123] Donald B. Rubin. "Comment: Neyman (1923) and causal inference in experiments and observational studies." In: *Statistical Science* 5.4 (1990), pp. 472–480.

[124] Guido W. Imbens. "Nonparametric estimation of average treatment effects under exogeneity: a review". In: *Review of Economics and Statistics* 86.1 (2004), pp. 4–29.

[125] Kosuke Imai. "Variance identification and efficiency analysis in randomized experiments under the matched-pair design". In: *Statistics in Medicine* 27 (2008), pp. 4857–4873.

[126] Laura B. Balzer et al. "Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching". In: *Statistics in Medicine* 35 (2016), pp. 3717–3732.

[127] K.L. Moore and M.J. van der Laan. "Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation". In: *Statistics in Medicine* 28.1 (2009), pp. 39–64. DOI: 10.1002/sim.3445.

[128] Michael Rosenblum and Mark van der Laan. "Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables". In: *The International Journal of Biostatistics* 6.1 (2010).

[129] Mark J. van der Laan, Laura Balzer, and Maya Petersen. "Adaptive Matching in Randomized Trials and Observational Studies". In: *Journal of Statistical Research* 46.2 (2012), pp. 1131–156.

[130] Laura Balzer, Maya Petersen, and Mark J. van der Laan. "Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation". In: *Statistics in Medicine* 34.6 (2015), pp. 999–1011.

[131] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. second. Wiley, 2011.

[132] Peng Li and David T. Redden. "Small Sample Performance of Bias-corrected Sandwich Estimators for Cluster-Randomized Trials with Binary Outcomes". In: *Statistics in Medicine* 34.2 (2015), pp. 281–296.

[133] Alisa J. Stephens, Eric J. Tchetgen Tchetgen, and Victor de Gruttola. "Locally efficient estimation of marginal treatment effects when outcomes are correlated: is the prize worth the chase?" In: *The International Journal of Biostatistics* 10.1 (2014), pp. 59–75.

[134] Cole Beck, Bo Lu, and Robert Greevy. *nbpMatching: Functions for Optimal Non-Bipartite Matching*. R package version 1.5.1. 2016. URL: https://CRAN.R-project.org/package=nbpMatching.

[135] S.R. Seaman, M. Pavlou, and A.J. Copas. "Review of methods for handling confounding by cluster and informative cluster size in clustered data". In: *Statistics in Medicine* 33 (2014), pp. 5371–5387.